А. Д. МЫШКИС ·

ЛЕКЦИИ ПО ВЫСШЕЙ МАТЕМАТИКЕ

A. D. MYŠKIS

INTRODUCTORY
MATHEMATICS
FOR ENGINEERS

Lectures
in
Higher
Mathematics

Translated from the Russian

by

V. M. VOLOSOV, D.Sc.

Preface

.

The present book is based on lectures given by the author over a number of years to students of various colleges studying engineering and physics. The book includes some optional material which can be skipped for the first reading. The corresponding items in the table of contents are marked by an asterisk.

In designing this course the author tried to select the most important mathematical facts and present them so that the reader could acquire the necessary mathematical conception and apply mathematics to other branches of science. Therefore in most cases the author did not give rigorous formal proofs of the theorems and intentionally simplified their statements referring the reader to characteristic particular cases and obvious examples. The rigorousness of a proof often fails to be fruitful and therefore it is usually ignored in practical applications. Some purely mathematical stipulations are made in the book only in the cases when they help the reader to avoid misconception in theory and application. Mathematical facts and objects which can be regarded as exceptional from the point of view of applied science are not even mentioned in the book. (For instance, when we speak about "all functions" we do not include the functions which are not Lebesgue measurable and even such functions as the everywhere discontinuous Dirichlet function and the like.) We tried to demonstrate the meaning of the basic mathematical concepts and to give a convincing explanation of the most important mathematical facts on the basis of intuitive notions. It is the author's belief that in applied mathematics an explanation of this kind should be regarded as a proof. Such an approach is characteristic of applied mathematics whose main purpose is to provide an adequate qualitative description of a phenomenon and obtain the numerical solution of the corresponding problem in the most economical manner without exerting unnecessary effort. This approach essentially differs from that of pure mathematics whose corner-stone is the logical consistency of all the considerations based

only on the concepts which have an exhaustive logical foundation. The author is sure that it is the aspects of applied mathematics that must determine the character of mathematical education of an engineer and physicist. (But of course a teacher of mathematics should have a good-command both of pure and applied mathematics.)

These ideas of the author concerning mathematical education (represented in greater detail in his article on applied mathematics published in the journal *Vestnik Vysshei Shkoly*, 1967, No. 4, pp. 74-80) are still difficult to realize consistently. Therefore the author will be grateful to the readers for any advice and criticism.

The book is composed in such a way that it is possible to use it both for studying in a college under the guidance of a teacher and for self-education. The subject matter of the book is divided into small sections so that the reader could study the material in suitable order and to any extent depending on the profession and the needs of the reader. It is also intended that the book can be used by students taking a correspondence course and by the readers who have some prerequisites in higher mathematics and want to perfect their knowledge by reading some chapters of the book. For this purpose we sometimes refer the reader to supplementary books (the bibliography is placed at the end of this course; the references are indicated by numbers in square brackets). We also supply the book with the name index, subject index and the list of symbols which enable the reader to find a desired definition, term or symbol.

In some colleges analytic geometry and linear algebra are studied as independent courses. The structure of the book facilitates such a separation: the fundamentals of analytic geometry and linear algebra are given in Chapters II, VI, VII, X and XI.

Some attention should be paid to the way of the numeration of the formulas and sections in the book. The sections entering into each chapter are numerated in succession beginning with the first number. In references inside each chapter we omit the number of the chapter. For instance, the expression "formula (2)" placed in the text of Chapter VI means "formula (2) of Chapter VI". But when formula (2) of Chapter VI is mentioned in some other chapter we write "formula (VI.2)". Similarly, "§ II.3" means "§ 3 of Chapter II" but we simply write "§ 3" when § 3 of Chapter II is referred to in this chapter; the expression "Sec. V.6" means "Sec. 6 of Chapter V" and so on.

Studying the theoretical material should be followed by solving problems and doing exercises. For this purpose we can recommend the well-known collections of problems [2], [4], [26] and [47]. But it should be noted that some divisions of applied mathematics are not treated to a sufficient extent in these collections and therefore it is advisable that a teacher of mathematics should add some interesting and instructive problems concerning these divisions.

The book can be of use to readers of various professions dealing with applications of mathematics in their current work. Modern applied mathematics contains, of course, many important special divisions which are not included in this book. The author intends to write another book devoted to some supplementary topics such as the theory of functions of a complex argument, variational calculus, mathematical physics, some special questions of the theory of ordinary differential equations and so on. ·

When preparing the book for the second edition the author considerably revised the text and added some new material including the chapter on the theory of probability*. Besides, the author has taken into account valuable advice and criticism received from many mathematicians, in particular from the members of the Moscow mathematical society where the book was discussed. Some sections of the book were written or revised under the influence of ideas and useful comments of L. M. Altshuler, Ya. B. Zeldovich and B. O. Solonouts. To all of them the author expresses his warmest gratitude.

A. D. Myškis

April 19, 1966

---

* This edition is the English translation of the second Russian edition of the book. The first Russian edition contained a chapter in which a brief review of basic equations of mathematical physics was given. The chapter was excluded from the second edition because of some changes in the syllabus of technical colleges. We have included the material of this chapter in this English edition as the Appendix at the end of the book. The present translation incorporates suggestions made by the author. — *Tr.*

Contents

(The starred items in the table of contents indicate those
sections which contain some optional material that may be
omitted for the first reading of the book.)

Introduction

1. The Subject of Mathematics. Numerical calculations are penetrating into the fields of work of physicists, chemists and engineers of various specialities. The modern development of science and engineering makes it necessary to deduce and apply still more complicated laws, to solve very complicated problems and perform extensive calculations.

All such calculations are based on mathematics, the science which treats of relations existing between spatial forms, quantities and magnitudes of the real world. All the basic notions of mathematics emerged and were developed in connection with the demands of natural sciences (physics, mechanics, astronomy etc.) and engineering. The appearance of more complicated problems led to the creation of more sophisticated mathematical methods of investigation (i.e. mathematical rules, techniques, formulas and the like) and, in particular, to the foundation of higher mathematics. It is therefore not accidental that the fundamentals of higher mathematics were created in the 17th and 18th centuries, i.e. at the beginning of an intensive development of industry, although some elements of higher mathematics appeared as early as antiquity in the works of the great Greek mathematician and mechanician Archimedes (287-212 B.C.).

Higher mathematics was founded in the works of the prominent French philosopher, physicist, mathematician and physiologist R. Descartes (1596-1650), the great English physicist, mechanician, astronomer and mathematician I. Newton (1642-1727), the great German mathematician and philosopher G. Leibniz (1646-1716), the great mathematician, mechanician and physicist L. Euler (1707-1783) and many other famous scientists. In their works different divisions of mathematics were created for investigating phenomena of nature and solving engineering problems. In mathematics, as in other sciences, practical work is the main source of scientific discoveries. Another important source is the need of mathematics

itself to systematize the facts discovered, to investigate their inter-relations and so on.

2. The Importance of Mathematics and Mathematical Education. Mathematics and, in particular, higher mathematics plays a very important role in modern natural science and engineering. Mathematics lies in the foundation of all divisions of physics, mechanics and many divisions of other natural sciences, engineering and some other branches of knowledge. Designing the construction of an airplane or a dam of a hydro-electric power station, investigating complicated processes involved in deformation of metals, propagation of radio-waves, diffusion of neutrons in an atomic reactor etc. cannot be performed without systematic application of mathematics.

The high level of development of computational methods in the USSR is one of the main factors that led to the triumphant achievements in launching the first artificial satellites of the Earth, space rockets and spacecraft. The creation of high-speed electronic computers and other mathematical automatic devices leads to further extension of the application of higher mathematics and facilitates the introduction of computational methods into many new fields. In particular, this is the case in such fields as economics, management and control of industry, elaborating optimal (i.e. the best) plans of capital investments or construction, transportation problems, controlling technological processes, dispatching and so on. The application of mathematical methods in these fields has already proved to be very effective and profitable. In recent years mathematics has been penetrating into such traditionally "non-mathematical" fields as biology, physiology, geography etc.

Therefore nowadays the requirements for mathematical education of an engineer are very high. An engineer must know the basic principles of higher mathematics and be able to apply them to concrete problems. Then mathematics will become a powerful tool in his hands. Besides, a great deal of scientific literature and many special technical subjects are saturated with mathematical techniques and formulas. Without sufficient knowledge of mathematics much effort is needed to understand all these formulas which may hinder the reader's work and mislead him. Mathematics also facilitates a better understanding of many questions related to other sciences (the theory of vibrations, mechanics of continuous media and so on).

3. Abstractness. Mathematics itself is not a technical subject and therefore a course in mathematics for engineers and scientists must not treat any special technical questions. Its aim is to provide the necessary mathematical education. Therefore a student may sometimes feel that the questions treated in a course of higher mathematics are too abstract. But the abstractness of mathematics is one

of its most essential features. This does not at all mean that mathematics has little to do with practical activities. On the contrary, it is the possibility to apply mathematics to various kinds of activities that makes its abstractness so important. For instance, in geometry we consider an "abstract" cylinder and find its volume. This immediately enables us to compute the volume of any concrete cylinder no matter whether it is a component of a mechanism or a column or a portion of space occupied by an electric field. Similarly, in higher mathematics we deduce some general abstract laws whose statements are not directly connected with a particular form of practical activity, a natural science or engineering, but the concrete applications and realizations of these laws (which are studied as examples in a mathematical course) are always related to various phenomena of the real world.

Thus, mathematics considers pure, ideal (schematized) forms, relations, processes etc. whose realization serves only as an approximation to reality. For instance, a real cylinder can never be a perfect cylinder from the mathematical point of view. Here we see the manifestation of a distinguishing feature characteristic of any kind of human cognition: when considering a real object or process we always select a number of basic properties from an infinite variety of properties of the object or process and investigate these most essential properties abstracting them from inessential ones. But it may sometimes happen that an assumption, hypothesis, that all the properties except those chosen as basic ones are inessential is not true and then we can arrive at a contradiction between our mathematical inferences and reality. Such a possibility must never be forgotten!

Because of the abstractness of forms and relations the logical consistency of inferences in mathematics is extremely important, more important than in other sciences, this being well known even from elementary mathematics. In higher mathematics too, all the assertions must be completely clear and logically justified so that it should be possible to regard them as objective laws adequate to reality. In mathematics, and particularly in higher mathematics, we also sometimes draw certain conclusions from experiment, observation and analogy but nevertheless such a situation is rarer in mathematics than in other sciences.

· There is a characteristic tendency in mathematics to deduce all the assertions from a few basic principles (called axioms). This is the so-called deductive method. But in our introductory course which is intended for those who are mainly interested in applications we shall not rigorously follow this method in all cases. The reader interested in theory may find some inferences in our course to be imperfect from the point of view of logic. If he wants to get a better understanding of some exceptions to general rules and to study higher

mathematics more thoroughly, he should study a course wri
for mathematicians, for instance, [14]. To consider the same ques
from different points of view it is advisable to take other cou
intended for technical colleges (for instance, [5], [37], [44] and
we particularly recommend book [5]).

4. **Characteristic Features of Higher Mathematics.** There is
distinct boundary between elementary and higher mathema
the division being conditional. These are not at all different scien
and the division is mainly accounted for by some historical rea
as elementary mathematics and higher mathematics were cre
in different historical epochs. But nevertheless we can point
some characteristic features of higher mathematics.

One of them is the universality, generality, of its methods
an example, let us take the problem of finding volumes of so
Elementary mathematics gives us different formulas for compu
the volumes of a prism, pyramid, cone, cylinder, sphere and s
other simple solids. Each formula is obtained on the basis of a
cial argument which is rather complicated in certain cases. Bu
higher mathematics we have general formulas expressing the vol
of any solid, the length of any curve, the area of any surface
the like. Take another example. Consider the problem of inves
ting the motion of a material point under the action of given fo
In elementary courses in physics (based on elementary mathema
methods) we study only uniform rectilinear motion, uniformly a
lerated rectilinear motion, uniformly decelerated rectilinear mo
and uniform circular motion, and it is rather difficult to investi
other types of motion by means of techniques of elementary ma
matics. But the methods of higher mathematics make it pos
to investigate any type of motion which can be encountere
practical problems.

There is another characteristic feature of higher mathem
(related to the above one). It is the systematic consideratio
variable quantities. When investigating various objects and
cesses by means of elementary mathematics we usually regard
important quantities as velocities, accelerations, densities, ma
forces etc. as being invariable, constant (and yet we attain
aim only in some simple cases). But if these quantities vary
siderably (as is often the case) we cannot regard them as being
stant. To solve such problems we usually apply higher mathema
There is a branch of higher mathematics (called *differential calc*
which is one of the earliest divisions of mathematics particul
intended for solving various problems connected with an inv
gation of the dependence of one quantity upon another. The q
tities and their interrelations can be of any nature (for insta
we can consider the relation between the acceleration, velc
and path length of a motion or between the density, mass and l

and the like). Therefore differential calculus deeply penetrates into various natural sciences and engineering.

The third characteristic feature of higher mathematics is the close relationship between its various divisions and the systematic unification of the computational, analytical (based on formulas) and geometric methods in contrast to elementary mathematics in which the connection between algebra and geometry is more or less accidental. In higher mathematics, the coordinate method reduces geometric problems to solving algebraic equations, graphs are used for representing relations between variable quantities, analytical methods of integral calculus are applied for computing areas and volumes of geometric figures and so on.

Some historical remarks will be given in due course in this book. But it is expedient to make some introductory notes here. The most important divisions of higher mathematics which now form the basis of the syllabus for engineers of many specialities were created in the 17th and 18th centuries. They include the coordinate method, differential and integral calculus etc. These divisions are represented in courses of higher mathematics for engineers mostly in the form they appeared after the works of Euler. L. Euler (a Swiss by birth) spent most of his life in Russia and died in Petersburg. Most of his works (473 out of 865) were published in Russia. His outstanding results in various divisions of mathematics, mechanics, physics and other sciences lie in the foundation of these divisions.

Mathematics was created by scientists of many countries. Among Russian mathematicians we should mention N. I. Lobachevsky (1792-1856), the creator of a non-Euclidean geometry. He also obtained some important results in other divisions of mathematics and initiated mathematical studies in Kazan. An intensive development of mathematics in Petersburg began with the works of the prominent mathematician Academician M. V. Ostrogradsky (1801-1862). The founder of the famous Petersburg mathematical school was the great Russian mathematician and mechanician Academician P. L. Chebyshev (1821-1894). He obtained many important results in various fields of mathematics and its applications to the theory of mechanisms, cartography etc.

After Chebyshev most prominent representatives of the Petersburg mathematical school were Academician A. A. Markov (1856-1922), a famous mathematician and the creator of the theory of random processes, and Academician A. M. Lyapunov (1857-1918), the founder of the theory of stability.

Since the second half of the 19th century mathematical investigations have been developing in Kiev, Moscow, Odessa, Kharkov, and other Russian towns.

5. **Mathematics in the Soviet Union.** In the Soviet Union there are many centres of mathematical research. Among prominent

Soviet mathematicians we should mention Academicians A. D. Aleksandrov, P. S. Aleksandrov, N. N. Bogolyubov, V. M. Glushkov, L. V. Kantorovich, M. V. Keldysh, A. N. Kolmogorov, M. A. Lavrentyev, Yu. V. Linnik, N. I. Muskhelishvili, P. S. Novikov, I. G. Petrovsky, L. S. Pontryagin, V. I. Smirnov, S. L. Sobolev, A. N. Tikhonov, I. N. Vekua, I. M. Vinogradov, and others.

Mathematics is being intensively developed both in old centres and in new ones in Baku, Erevan, Gorki, Lvov, Minsk, Novosibirsk, Rostov, Saratov, Sverdlovsk, Tashkent, Tbilisi, Vilnyus, Voronezh, and other towns.

The role of mathematics in other sciences, industry and engineering has considerably increased. Many mathematicians work out new theoretical problems of other branches of knowledge connected with applications of mathematics. At the same time many physicists, mechanicians and engineers take part in the development and applications of those divisions of mathematics which are related to their fields of work. As examples of fruitful unification of mathematics and its applications we can mention the works of the great Russian scientist and one of the founders of modern flight mechanics and hydromechanics N. E. Zhukovsky (1847-1921), the prominent Russian scientist, mathematician, mechanician and naval architect Academician A. N. Krylov (1863-1945), the prominent Soviet scientist in the fields of theoretical mechanics, aerodynamics and hydromechanics Academician S. A. Chaplygin (1869-1942) and others.

There is no doubt that development of mathematical education will further increase the role of mathematics in our life and yield fruitful results.

Variables and Functions

§ 1. Quantities

1. Concept of a Quantity. It is difficult to give a strict definition of a quantity since the notion is extremely general and universal. Masses, pressures, charges, different kinds of work, lengths and volumes are examples of quantities. It will be sufficient for our further aim to regard as a quantity *everything that is expressible in certain units and completely characterized by its numerical value.* For instance, masses are measured in grams or kilograms and the like. We can say that the area of a circle is a quantity since it is completely characterized by its numerical value (for example, 5, π etc.) if we measure it in certain units, e.g. in square centimetres. The circle itself regarded as a geometric figure is of course not a quantity because it is characterized by a certain geometric form which cannot be expressed numerically.

Many notions which were originally understood only in a qualitative aspect have been recently "advanced" and transferred to the class of quantities (for instance, such notions as effectiveness. information and even likelihood). Every change of this kind is a great event since it enables us to apply quantitative mathematical methods to investigating the corresponding notions and this usually turns out to be very effective.

2. Dimensions of Quantities. *A unit measure which is used for expressing a quantity is called the dimension of the quantity.* For instance, the gram or the kilogram usually serves as the dimension of mass. The dimension of area is the square centimetre or the square metre and so on. A dimension is denoted by square brackets. For instance, if M is a mass and S is an area then $[M] = $ kg (the kilogram) and $[S] = $ m^2 (the square metre) in the international system of units.

Usually the units of some quantities are regarded as fundamental units whereas the units of all other quantities are derived units expressed in terms of the fundamental ones. For instance, the units of length (m), of mass (kg) and of time (sec) are the fundamental

units in the international system of units (SI), and the unit of velocity (m/sec) or of force (kg·m/sec²) is expressed in terms of the fundamental units.

We can add together and subtract only quantities of the same dimension, the dimension of a sum being that of the summands. It is permissible to multiply or divide quantities of arbitrary dimensions. The multiplication or division of quantities yields, respectively, the multiplication or division of their dimensions.

We also consider **dimensionless** ("abstract") quantities. For instance, the ratio of two quantities of the same dimension is dimensionless. The numerical value of the ratio of a quantity to the chosen unit measure is also dimensionless. For example, the numerical value of the mass of 5 kg is the "dimensionless mass" 5. We can also obtain a dimensionless mass if we take the ratio of the mass to a certain mass which is characteristic of the process in question (such a mass is supposed to be well known, and we choose it as a standard to compare with). Dimensionless length, time etc. are introduced in like manner.

In mathematics we usually regard quantities as dimensionless. Finally, a dimensionless quantity is completely characterized by its numerical value, and its "unit measure" is the number 1.

3. Constants and Variables. *A quantity entering into an investigation can take on either different values or only one fixed value.* In the first case we call the quantity a **variable quantity** or, in short, a **variable**, and in the second case we call the quantity a **constant** (a **constant quantity**). Suppose we consider the water in a basin. The water pressure measured at different points of the basin is a variable since it varies and is different at different points. At the same time the water density can be regarded as a constant since it takes on one and the same value (with a sufficient degree of accuracy) at different points. As another example let us consider the process of compressing a given mass of a gas while the temperature is kept constant. Then the pressure and the volume are variables whereas the mass and the temperature are constants. But it should be noted that in a real process the last two quantities inevitably vary a little. Hence, we can schematize the process and conditionally regard the mass and the temperature as constants only in case their real variations are of no importance for our investigation. And in many other cases the constancy of some quantities should be understood in a conditional sense. We must never forget it since, if we regard a quantity as a constant in a process in which the variations of the quantity, small though they may be, are essential for the investigation, we may arrive at wrong conclusions and our schematized model will not apply.

. It may happen that a quantity which is constant in a certain treatment of a phenomenon takes on a different value or even becomes

a variable under some other (though similar) circumstances. The constant quantities of this kind are called the **parameters** of the process; they are the characteristics of the process. For example, the mass and the temperature of a gas are the parameters of the process of isothermal compression. When we deal with an electric-light bulb we take into account such parameters as the resistance, the supply voltage the bulb is designed for and the power consumption. Even in this case there are some other parameters which may also be taken into account (for instance, the sizes of the bulb) but usually we do not regard these parameters as basic ones. Generally, in all cases it is very important to choose the basic, the most significant parameters among various parameters characterizing an object.

4. **Number Scale. Slide Rule.** Quantities can be represented visually by means of a **number scale**. For this purpose we usually take a **rectilinear axis with a uniform scale**. To construct a number scale we choose a straight line and a point on the line which serves as the origin (the origin is usually designated by the letter O). We choose one of the directions on the straight line as the positive direction and take a certain line segment as a unit of length (the positive direction is indicated by an arrow; see Fig. 1). Setting off
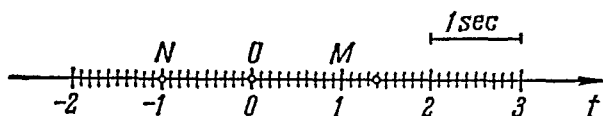
Fig. 1

the unit segment from the origin in both directions and repeating the procedure infinitely we obtain all the points which correspond to the integral values of the quantity. Between the "integer points" there are points representing fractional values, both rational $\left(\text{such as } \frac{1}{2}, \ -2.03 \text{ etc.}\right)$ and irrational (that is fractional numbers that are not rational, e.g. $\frac{\sqrt{2}+1}{3}$, $-\pi$ and the like). In case we have a dimensional quantity the segment chosen as the length unit also acquires the corresponding dimension. For example, the numerical values of time t depicted in Fig. 1 are expressed in seconds; we also see the points N ($t = -1$ sec), O ($t = 0$ sec) and M ($t = 1.37$ sec) there.

To each value of the quantity there corresponds a certain point on the number scale and, conversely, each point on the number scale corresponds to a certain value of the quantity. Besides, there is a one-to-one correspondence between the values of the quantity and the

points, that is each point corresponds only to one value and vice versa. (Here and further we consider real quantities, that is quantities which take on only real numerical values; complex quantities will be treated in Sec. VIII.1.) On the basis of these properties we often identify the values of a quantity with the corresponding points; we simply say "the point $t = 1.37$ sec" and the like.

If a quantity is variable it is represented by a point which can occupy different positions on the axis (on the number scale). For example, such a point can move along the axis as time passes. In

100   100.01 100.02 100.03 100.04 100.05 l cm
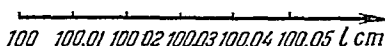
Fig. 2

case the quantity is constant the corresponding point occupies a fixed position and does not move. A point which represents a variable is called a **variable point** (**moving point, current point**).

In practical applications we try to choose the origin and the unit of length in such a way that the range of variations of the quantity should be represented in the most suitable manner. The origin
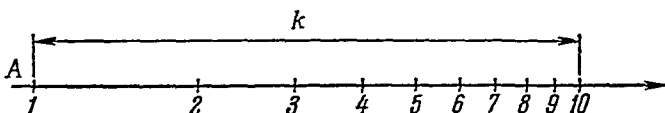
Fig. 3

itself is sometimes not depicted because we can draw only a part of the infinite axis. For example, Fig. 2 represents the number scale on which the values of the length of a rod subjected to thermal expansion are shown.

It is sometimes convenient to use scales which are non-uniform. For instance, **logarithmic scales** are often of use (see Fig. 3). A number $n > 1$ is represented on such a scale by a point which is obtained by drawing the line segment of length $k \log n$ (where k is a factor of proportionality suitably chosen) in the positive direction from a point A. Positive numbers $n < 1$ are obtained on the logarithmic scale by drawing the segment $k |\log n|$ in the negative direction from A because for such n we have $\log n < 0$.

A logarithmic scale is, in particular, utilized in the construction of a **slide rule**. The instrument consists of a ruler and a slide which are graduated with similar logarithmic scales. To understand the principle of a slide rule let us suppose that the scales are shifted

with respect to each other (see Fig. 4) so that two points a and b on the lower scale coincide with the corresponding points a' and b' on the upper scale. Then we have

$$k \log b - k \log a = k \log b' - k \log a'$$

since the lengths of the shaded line segments are equal. Now after some simple transformations we obtain (check it up!)

$$\frac{b}{a} = \frac{b'}{a'} \qquad (1)$$

Three of the values a, b, a' and b' being given, we can read the fourth value on the slide rule. This fourth value will satisfy relation (1). If, for example, we put $a' = 1$ then $b = ab'$ or $a = \frac{b}{b'}$. Consequently, to determine the product of two given numbers a and b' we must make the point 1 on the slide coincide with the point a on the ruler
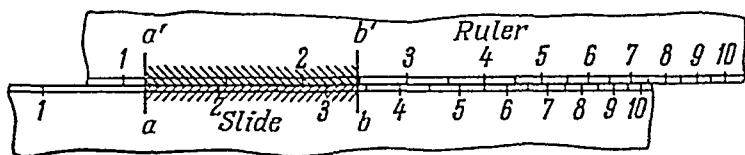
Fig. 4

and then read the value of the product which is indicated on the ruler by the point b' of the slide. (Think how to find the quotient of two given numbers.) It is sometimes convenient to put $b' = 10$ instead of $a' = 1$ and to move the slide not to the right but to the left with respect to the ruler. The slide rule was invented in the 17th century. It is widely used now and facilitates the work of many technicians, engineers, physicists etc. Supplementary scales on the slide rule make it possible to perform various additional operations including extracting roots, taking logarithms, raising, solving equations of different types and so forth. There is a number of handbooks on using the slide rule, for example, [36] and [43] to which we refer the reader.

Some curvilinear scales are also of use in certain cases (for example, see Sec. IX.1). But in our course we shall usually use rectilinear axes and uniform scales for representing quantities unless the contrary is explicitly stated.

5. **Characteristics of Variables.** A variable which takes on all the numerical values or all the values lying between some limits is called **continuous.** On the contrary, a variable which assumes certain "separated" values is called **discrete.**

The set of all numerical values which may be assumed by a variable is called the **range of the variable.**

Now we introduce the notion of an interval which is of use for characterizing ranges of some types of variables.

A **finite (bounded) interval** is the set of all numbers contained between two given numbers a and b. The numbers a and b are called the **end-points of the interval.** The end-points a and b may or may not be included into the interval and this fact should be sometimes indicated. Respectively, in the first case we call the interval **closed** (i.e. when $a \leqslant x \leqslant b$ and the end-points are thus included) and denote it as $[a, b]$ and in the second case we say that the interval is **open** (i.e. $a < x < b$ and the end-points are excluded) and denote it by (a, b). Finite intervals are represented by line segments on the number scale.

There are also **unbounded (infinite) intervals** for which a or b or both a and b may be infinite. For example, if a variable x assumes all possible values greater than some constant number a the range of the variable is described by the inequalities $a < x < \infty$. This is an example of an infinite interval; it has no finite right end-point, of course, but in such a case we say, conditionally, that the right end-point is at infinity. An interval of this kind is also said to have **no upper bound** since in case a variable may increase unlimitedly we usually interpret the variable as "rising up". The notion of a **lower bound** is understood in just like manner. The collection of all real numbers is an interval with neither lower nor upper bound (that is, geometrically, the whole number scale).

The range of a continuous variable is an interval or a collection of some number of intervals. For example, if a triangle ABC is deformed in all possible ways the corresponding angle A is a continuous variable whose range is the interval $0 < \angle A < \pi$ (in case the numerical values of the angle are expressed in radians). At the same time the area S of the triangle has the interval $0 < S < \infty$ as its range (of course, here we also mean that the numerical values of the area are measured in certain units but we are not going to mention details of this kind in all cases in future). The range of a discrete variable is a set (finite or infinite) of separate real numbers. We can also say, in the geometric sense, that such a range consists of separate points (but not of entire intervals). For example, let an index assume the values $1, 2, \ldots, n$. Then it is a discrete variable.

If a variable changes in a certain process in such a way that its numerical values vary only in one direction, that is they either increase or decrease, it is called **monotonic.** The point representing a monotonic variable on a number scale moves in one direction.

It is inconvenient to consider constant quantities apart from variables and therefore we can regard a constant quantity as a special case of a variable, i.e. a variable which all the time assumes

only one fixed value (the same idea is used in mechanics when the state of rest is regarded as a special case of motion). The range of a constant consists of only one point.

We say that a variable changing in a certain process has an upper bound if all the time it remains smaller than a constant (such a constant is called an upper bound of the variable; it is clear that a variable having an upper bound has in fact an infinitude of upper bounds because every constant greater than a given upper bound of the variable can serve as a new upper bound). We likewise define the notion of a lower bound of a variable. Of course, a variable may not have an upper or a lower bound (or either of them). If a
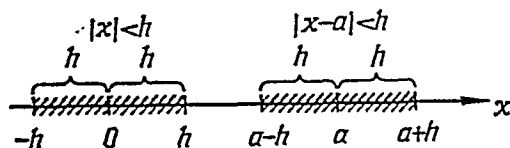
Fig. 5

variable has both an upper bound and a lower bound it is simply called a **bounded variable**. Variables having upper bounds (lower bounds) are called **bounded above (bounded below)**.

When investigating different quantities we often use the notion of the **absolute value** of a quantity. As is well known from elementary mathematical courses, the notion is defined in the following way:

$$|a| = a \text{ if } a \geqslant 0 \quad \text{and} \quad |a| = -a \text{ if } a < 0$$

For instance, $|5| = 5$, $|0| = 0$ and $|-5| = 5$ [that is $|-5| = -(-5) = 5$].

Absolute values possess the following simple properties:

1. $|a + b| \leqslant |a| + |b|$. The inequality is strict in case a and b have opposite signs and it turns into the equality if otherwise.

2. For any a and b we have

$$|ab| = |a| \cdot |b| \quad \text{and} \quad \sqrt{a^2} = |a|$$

The significance of the last formula is sometimes underestimated in elementary mathematical courses and this may be the cause of different errors and false conclusions.

The quantity $|a - b| = |b - a|$ is equal to the distance between the points a and b lying on the number scale. The inequality $|x| < h$ ($h > 0$) defines the interval $-h < x < h$, and the inequality $|x - a| < h$ defines the interval $-h < x - a < h$, i.e. $a - h < x < a + h$. An interval of the form $a - h < x < a + h$ is called an h-**neighbourhood** of the point a. The intervals are shaded in Fig. 5.

§ 2. *Approximate Values of Quantities*

6. The Notion of an Approximate Value. It is usually impossible to speak about the absolutely precise value of a physical quantity. For example, we can never determine the exact value of the length of a real object. This is so not only because our measurements are imperfect but also because of a complex form of the body which makes it impossible to indicate exactly the points between which the length should be measured. If we recall that the object consists of molecules which are in permanent motion we see that the situation becomes still more complicated. Moreover, there is a vast majority of cases when the determination of a length with a great accuracy is inexpedient and senseless even when the modern level of measurement techniques makes such an accuracy attainable. For instance, if we have to design or measure a dwelling-house it would be obviously senseless to determine the sizes of the building with the accuracy to within 0.01 mm. The same can be said about masses, pressures etc. The numerical values of almost all quantities in physics and engineering (for example, the values of all continuous variables) are therefore approximate.

Mathematical operations on approximate values of quantities are called **approximate calculations**. There exists a special branch of science devoted to approximate calculations and we shall study some of its rules later on. A. N. Krylov (1863-1945) was one of the initiators of developing approximate calculations in the USSR. His book [28] (the first edition was in 1911) still retains its significance.

The appropriate choice of a degree of accuracy for calculations, measurements or for manufacturing machine elements is a very important operation. When making such a choice one should take into account a great many factors, i.e. our requirements, technical means, economy etc.

7. Errors. Let A be the exact value and a an approximate value of a quantity. Then the **error**, that is the deviation of the approximate value from the exact one, is equal to $A - a$. It may be positive or negative. As a rule, we do not know the error exactly since the exact value A is unknown. Therefore we usually consider the **limiting errors** α_1 and α_2 which form an interval containing the true error:

$$\alpha_1 < A - a < \alpha_2, \quad \text{i.e.} \quad a + \alpha_1 < A < a + \alpha_2$$

Thus the value of the quantity A is estimated from two sides. For instance, the formula of the length $L = 9^{+0.2}_{-0.1}$ mm means that the true value of the length lies between $9 - 0.1 = 8.9$ mm and $9 + 0.2 = 9.2$ mm.

It is sometimes inconvenient to consider two limiting errors and therefore we often indicate the **maximum absolute error** α, that is

a value which exceeds the absolute value of the error:

$$|A - a| < \alpha, \quad \text{i.e.} \quad -\alpha < A - a < \alpha \quad \text{or}$$
$$a - \alpha < A < a + \alpha$$

For example, suppose that the measurement of a length l results in the value 137 cm and that we can guarantee the accuracy of 0.5 cm. This means that we have $\alpha = 0.5$ cm and 136.5 cm $< l <$ < 137.5 cm. Therefore we can write $l = (137 \pm 0.5)$ cm.

The maximum absolute error of a measurement does not characterize it completely. For instance, if we are told that the maximum absolute error is equal to 1 cm we do not yet know whether this is a great error or not. Indeed, for example, if it were the length of a whale or of a beetle our judgment would vary respectively.

The quality of a measurement is better characterized by its **maximum relative error** δ which is calculated by the formula

$$\delta = \frac{\alpha}{|a|}$$

The maximum relative error is dimensionless and we often express it in per cent. The value of a relative error is usually rounded for the sake of simplicity. For instance, the relative error in per cent for the example of measuring the length l is equal to $\frac{0.5 \times 100}{137} =$ $= 0.36 \approx 0.4$, i.e. we can say that the maximum relative error of the measurement is equal to 0.4% (or, rounding, to 0.5%).

The accuracy of the order of 1% or even of 10% is sufficient for many approximate calculations. On the other hand, the precise measurement of the frequency of electromagnetic vibrations which creates the basis of performing the automatic control of spacecraft is carried out by means of a crystal or an atomic clock whose error in time observation is about 10^{-4} sec a day (calculate the corresponding maximum relative error!).

8. Writing Approximate Numbers. It is desirable to write an approximate number, i.e. an approximate value of a quantity, in a form which indicates the degree of accuracy. Therefore an approximate number is usually written in such a way that *all the decimal digits except the last one are correct. The admissible error for the last decimal digit must not exceed unity* (by the way, if the error is a little greater one usually admits it). For instance, if we write the value $R = 1.35$ Ω of a resistance we mean that $\alpha_R = 0.01$ Ω, that is in fact 1.34 $\Omega < R < 1.36$ Ω. There is a great difference between the formulas $R = 1.35$ Ω and $R = 1.3500$ Ω since the former indicates that the corresponding calculations were carried out with a possible error of 0.01 Ω whereas the latter expresses the result accurate to 0.0001 Ω. If the result of certain calculations is $R = 2.377$ Ω but the third decimal digit may be incorrect, or

if we are not interested in the fourth decimal digit, we should **round off** the result and write $R = 2.38\ \Omega$.

The number of decimal places to the right of the decimal point indicates the maximum absolute error. The total number of correct decimal digits (which does not include zeros standing to the left of the first nonzero decimal digit) indicates the maximum relative error. For instance, the numbers 2.57; 1.7100; 0.015 and 0.00210 have, respectively, 3, 5, 2 and 3 correct decimal digits. The greater the number of correct decimal digits, the smaller the maximum relative error.

One should avoid writing expressions of the form $M = 1800$ g since such a form often does not indicate the real accuracy of measurements or calculations. If we suspect that the second decimal digit may be incorrect we must write $M = 1.8 \times 10^3$ g, and if the fourth decimal digit is doubtful we must write $M = 1.800 \times 10^3$ g. Strictly speaking, the formula $M = 1800$ g means that the maximum absolute error is equal to 1 g. When the rules of writing approximate numbers are not followed we often encounter various misunderstandings.

9. Addition and Subtraction of Approximate Numbers. Let us take an example. Suppose we have weighed a bottle and its cork. Let their masses be, respectively, $M = 323.1$ g and $m = 5.722$ g (this indicates that the scales taken for weighing the cork are more precise). It would be wrong to calculate the total weight of the bottle together with the cork as

$$
\begin{aligned}
M &= 323.1 \\
m &= 5.722 \\
\hline
M + m &= 328.822\ \text{g}
\end{aligned}
$$

Actually, the weight of the bottle is found with an accuracy of 0.1 g and therefore the hundredths and the thousandths entering into the result are not only unnecessary but even misleading: judging by the answer one might think that the value of $M + m$ is accurate to 0.001 g which is wrong. To perform the addition correctly we must therefore round off m to 0.1, that is we must calculate in the following way:

$$
\begin{aligned}
M &= 323.1 \\
m &= 5.7 \\
\hline
M + m &= 328.8\ \text{g}
\end{aligned}
$$

Of course, we should obtain the same result if we rounded the former result but such an operation involves the unnecessary calculations which should have been avoided. Thus, the number of decimal digits entering into a sum must be the same as in the summand with the greatest absolute error.

If there are many summands the round-off errors may add up and this may result in a great error in determining the sum (some

kind of systematic "short measure"). There is a **rule of a reserve decimal digit** which is recommended for such cases: the calculations are carried out to an extra decimal digit and then we round off the result discarding the reserve decimal digit after the sum has been calculated.

For example, let it be required to find the sum

$$K = 132.7 + 1.274 + 0.06321 + 20.96 + 46.1521$$

The first summand has the largest absolute error which is equal to 0.1. Therefore we round off all the other summands to 0.01:

$$132.7 + 1.27 + 0.06 + 20.96 + 46.15 = 201.14$$

Now, rounding, we find $K = 201.1$. If we did not use the rule the result would be less accurate:

$$K = 132.7 + 1.3 + 0.1 + 21.0 + 46.2 = 201.3$$

Another example. Suppose it is necessary to find the sum $N = \sqrt{5} + \sqrt{6} + \sqrt{7} + \sqrt{8}$ with an accuracy of 0.01. The integers under the radical signs are regarded as quite exact. Using the rule of a reserve decimal digit we take from a table the values of the roots accurate to 0.001:

$$2.236 + 2.449 + 2.646 + 2.828 = 10.159$$

Thus, $N = 10.16$.

If the number of summands is very large, say several hundreds, it is advisable to use two reserve decimal digits.

When we add several summands given with the same number of decimal digits to the right of the decimal point we must take into account that the maximum absolute error of the sum can be greater than those of the summands. It is therefore expedient to round the result to the preceding decimal place. For example, let

$$L = 1.38 + 8.71 + 4.48 + 11.96 + 7.33$$

Adding we get $L = 33.86$. But the last decimal digit is likely to be incorrect and therefore we should write the answer in the form $L = 33.9$.

The maximum absolute error of a sum or of a difference is equal to the sum of the maximum absolute errors of the operands. For instance, if we have two quantities determined with an accuracy of 0.1 then it is easy to understand that the sum and the difference of the quantities are determined with an accuracy of 0.2 because the errors may add up. If there are many summands it is unlikely that all the errors would add up. In such cases one should use the methods of the theory of probabilities (see Sec. XVIII.15) in order to estimate the error of the sum. These methods imply that we should round the sum

discarding one decimal digit (as it was done above in calculating L) beginning with, approximately, five summands and two decimal digits approximately with 500 summands.

The rules of subtraction are essentially the same as the rules of addition of approximate numbers. At the same time we should take into account that when subtracting approximate numbers which are close to each other we may get a considerable increase of the relative error. For instance, let it be necessary to calculate $P = 327.48 - 326.91$. The minuend and the subtrahend have $\alpha = 0.01$ and therefore $\delta \approx \frac{0.01}{300} 100\% = 0.003\%$. But the maximum absolute error of the difference $P = 0.57$ is equal to 0.02 and hence its maximum relative error is $\delta_P = \frac{0.02}{0.57} 100\% = 3.5\%$. The relative error has thus increased 1000 times!

Therefore one must try to avoid calculating the difference of two close numbers. In such a case we should transform the corresponding expressions in an appropriate way in order to find the difference without actually carrying out such a subtraction: one must not try to determine the weight of one's hat by weighing oneself first with the hat on and then without it!

When we deal with formulas containing differences of this kind which can noticeably affect the accuracy of calculations we should eliminate the differences by transforming the expressions. For example, calculating an expression of the form $Q = a - \sqrt{a^2 - b^2}$ ($a > 0$, $b > 0$) where b is several times smaller than a (and therefore $\sqrt{a^2 - b^2} \approx \sqrt{a^2} = a$) we can transform the expression in the following way:

$$Q = \frac{(a - \sqrt{a^2 - b^2})(a + \sqrt{a^2 - b^2})}{a + \sqrt{a^2 - b^2}} = \frac{b^2}{a + \sqrt{a^2 - b^2}}$$

The last expression no longer contains the undesirable difference.

10. **Multiplication and Division of Approximate Numbers. General Remarks.** Let us begin with an example. Suppose it is necessary to determine the area S of a rectangle with the sides $a = 5.2$ cm and $b = 43.1$ cm. It would be wrong to give the answer $S = 5.2 \times 43.1 = 224.12$ cm².

In fact, a is contained between 5.1 and 5.3, and b between 43.0 and 43.2. Thus, the area is contained between

$$S_1 = 5.1 \times 43.0 = 219.3 \text{ cm}^2 \quad \text{and}$$
$$S_2 = 5.3 \times 43.2 = 228.96 \text{ cm}^2$$

We see that all the decimal digits beginning with the second one in the above value of S may be incorrect and therefore they may only

lead to misconceptions. In this case the correct answer we must give is $S = 2.2 \times 10^2$ cm^2.

By the way, we note that the calculation of S_1 and S_2 demonstrates the way that can be followed in estimating the results in other problems.

Thus we see that in multiplying two numbers with two and three correct decimal digits we must retain two decimal digits in the answer. The same rule holds for the general case of multiplication of approximate numbers and also for their division: the number of correct decimal digits in the result must be equal to the smallest of the numbers of correct decimal digits in the factors (or in the divident and the divisor in the case of division). The reason for this general rule is that, in the first place, the operations of multiplication and division performed on approximate numbers yield the addition of the corresponding maximum relative errors (this will be shown in Sec. IX.11) and, in the second place, the number of correct decimal digits and the maximum relative error indicate similar qualities connected with the degree of relative accuracy.

In the example of calculating S the maximum relative error of b is considerably smaller than that of a and therefore $\delta_S = \delta_a + \delta_b \approx \delta_a$, that is S has the same number of correct decimal digits as a.

If the factors entering into a product are given with different numbers of correct decimal digits we must round the numbers before multiplying them and retain one reserve decimal digit which is discarded after the operation is performed. In case the factors have the same number of correct decimal digits but there are many factors (for instance, more than four) it is advisable to reduce the number of correct digits in the product by one.

As an example, let us take the formula $Q = 0.24 I^2 R t$ which is applied to calculating heat generated by an electric current. In this case the answer cannot have more than two correct decimal digits because the coefficient 0.24 has only two correct digits. Therefore there is no sense in taking I, R and t with more than three correct decimal digits (moreover, the third digit is taken only as a reserve digit). If a more accurate value of Q is desirable we must first of all specify the value of the coefficient.

It should be noted that absolutely exact factors do not affect the choice of the number of correct decimal digits in a product. For instance, the coefficient 2 entering into the formula $L = 2\pi r$ of the circumference of a circle is absolutely exact (we can write it as 2.0 or 2.00 etc.) and therefore the accuracy of calculations depends only on the number of correct decimal digits to which π and r are computed.

Let us take an example involving all the above rules. Let $D = 11.3^2 \times 5.4 + 0.381 \times 9.1 + 7.43 \times 21.1$. In order to estimate

the magnitude of the summands we calculate them rounding to one correct decimal digit. Thus we get 500. 3.6 and 140. Hence the sum is of the order of several hundreds. The factor 5.4 entering into the first summand (which is the largest one) has only two correct decimal digits and thus the whole result must have two correct digits. Now, according to the rule of a reserve decimal digit, we must calculate to within unity and then round off the result to the nearest ten. Thus we obtain $D = 690 + 3 + 157 = 850$, i.e. $D = 8.5 \times 10^2$.

Calculations with unnecessary digits are not only useless but even misleading because they may give the illusion of an accuracy greater than that we actually have.

The choice of a degree of accuracy of approximate quantities for performing mathematical operations on them is made in accordance with a general principle which states that all the degrees of accuracy which we choose must be coherent to each other at every stage of our calculations. This means that none of the degrees must be too great or too low.

We shall take an example to illustrate the principle. Suppose we have to calculate the area of a rectangle by the formula $S = ab$. Let a be measured or calculated to three correct decimal digits. Then we must take b also with three correct digits because the fourth decimal digit of b would be useless whereas if we determined b only with two correct digits the efforts applied to finding the third digit of a would be futile. Therefore when we calculate a product it is convenient to take the factors (at least those factors which are difficult to determine) with the same number of correct decimal digits. Similarly, the summands entering into a sum must be taken with the same number of decimal digits to the right of the decimal point.

Here we give an example. Let the expression $M = ab + cd$ be calculated and let it be known that $a \approx 30$, $b \approx 6$, $c \approx 0.1$ and $d \approx 40$. Suppose that a is taken with three correct decimal digits. What number of correct digits should be chosen for b, c and d? It is clear that we must take three correct decimal digits for b according to the accuracy of a. Further, we have $ab \approx 180$ and $cd \approx 4$. This implies that for calculating M with three correct digits (the accuracy of a makes it impossible to obtain M with more than three correct digits) it is sufficient to determine c and d with only one correct decimal digit. If it is not too difficult the accuracies of b, c and d should be increased by one decimal digit but the extra digit is only a reserve one.

When performing practical calculations we often face a problem which is in some sense inverse to the above problem. The degree of accuracy of a desired result is sometimes set beforehand according to some prerequisites and then it is required to determine the necessary

degrees of accuracy of the quantities involved into the calculations (and the accuracy of the calculations). Some of the quantities may be obtained as a result of an experiment and therefore our discussion also applies to the determination of a desirable precision of an experiment. The solution of the inverse problem is based on the rules of approximate calculations we have studied here. For example, suppose we have to calculate the total surface area of a circular cylinder by the formula $S = \pi\,(DH + \frac{D^2}{2})$. Let. it be approximately known that $D \approx 20$ cm and $H \approx 2$ cm. Then $S \approx 700$ cm² (check it up!). Now turning to the inverse problem and reasoning as in the preceding paragraph we see that if, for instance, we want to have the result with three correct decimal digits, i.e. with an accuracy of 1 cm², then π and D should be taken with three correct decimal digits and H with two correct digits. Thus, measuring D and H we must attain the accuracy of 1 mm. It is better to calculate with a reserve decimal digit, and π should also be taken with a reserve digit. But if we wanted to have more accurate values of D and H we would have to perfect our measuring instruments.

The rules of determining degrees of accuracy for more complicated formulas will be given in Secs. IV.10 and IX.11.

§ 3. Functions and Graphs

11. Functional Relation. When investigating a phenomenon or a problem we often deal with several variables which are interrelated so that a change of one of the variables affects the values of the others. Then we say that there is a **functional relation** between the variables. For example, suppose a mass of a gas is kept under changing conditions. Then there is a functional relation between the volume V, the temperature T and the pressure p of the gas because, as is well known from physics, the quantities are interrelated. We also have a functional relation between the area of a circle and its radius, between the distance passed over in a process of motion and the time taken and so forth.

Usually it is possible to pick out certain variables from a number of interrelated quantities such that the values of the variables can be taken arbitrarily whereas the values of the other quantities are determined by the values of the variables entering into the first group. The variables of the first type are called **independent variables (or arguments)** and the variables of the second type are called **dependent variables (or functions)**. As an example let us consider the relationship between the area S of a circle and the length R of its radius. It is natural to regard R as an independent variable and choose its values arbitrarily; then the area computed by the formula $S = \pi R^2$ is a dependent variable in this functional relation.

In the example of the mass of gas we could have taken V and T as independent variables. Then the variable p (the pressure) would have been regarded as a dependent variable.

A rule (a law) according to which to the values of independent variables there correspond the values of a dependent variable is called a function. Thus, every time there is a law of correspondence between the values of variables we say that there is a functional relation. The concept of a function is one of the most important mathematical notions.

By the way, the term "function" is sometimes used in a different sense. As it has been mentioned, independent variables are called arguments and a dependent variable itself is called a function. Such a twofold sense of the term does not, however, lead to any misunderstandings.

It should be noted that when we have a functional relation between variables the distinction between the independent variables and the dependent ones is sometimes conditional. For instance, in the example of the mass of gas we could have taken T and p as independent variables and V as a function. We can easily construct the scheme of an experiment in which T and p can be varied arbitrarily whereas V depends on T and p. Of course, the choice of variables which are regarded as independent quantities may be important in some cases. The choice should be made in a natural and convenient way in accordance with the circumstances.

Functions may depend on one argument (as in the example of the area of a circle) or on two or more arguments. In the first two chapters of our course we shall consider (almost without exceptions) functions of one independent variable.

We must note that when we regard a quantity y as a function of an independent variable x we do not necessarily suppose that there is a meaningful causal relationship between the variables. It is quite sufficient if there exists a rule which attributes a certain value of y to each x even if we do not know this law of correspondence. For example, the temperature θ at a point in space can be regarded as a function of time t since it is clear that we always have a certain temperature at the point at each moment t, that is to the values of t there correspond the values of θ, although the variations of θ cannot be simply accounted for by the changes of t since in reality these variations are determined by some complicated physical laws.

12. Notation. If a variable y is a function of a variable x we usually write $y = f(x)$ (this is read as "y is equal to f of x") where f is the sign of a function. If we make x assume certain **particular (concrete) values** the function will assume its particular values.

For instance, let $y = f(x)$ be of the form $y = x^2$. Then $y = 4$ for $x = 2$, $y = 0.36$ for $x = -0.6$ etc. This can be written as $f(2) = 4$, $f(-0.6) = 0.36$ and so on, or as $y|_{x=2} = 4$, $y|_{x=-0.6} =$

$= 0.36$ etc. The vertical lines in the last expressions are the signs of substitution which mean that we substitute the values of x for the argument.

The notation $y = f(x)$ is used when the concrete expression of a function is too complicated or when we do not know the expression. It is also used for formulating general rules and properties of all functions or of many concrete functions. For example, the formula $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$ which is well known from algebra is written in letters. Here the letters a and b are not concrete numbers but they can be replaced by any numbers.

If we consider several functions simultaneously we can use, besides f, any other letters: F, φ, Φ etc. We can also introduce different subscripts, superscripts, and other indices: f_1, f_2, F_7^s etc. At the same time when we consider different problems we can denote different functions by the same letter f. We remind the reader that we have a similar situation in algebra: a letter, say the letter a, may denote different quantities in different problems, but we must not denote by a different quantities entering into one and the same problem. On the other hand, different quantities may sometimes be connected by one and the same functional relation. In such a case we can use one and the same letter f because f designates the law of dependence of one quantity upon another and is irrelevant to the way the quantities are denoted. For example, if $y = x^3$, $z = u^5$ and $v = t^3$ then we can write $y = f(x)$, $z = \varphi(u)$ and $v = f(t)$. In this case the sign f indicates raising to the third power whereas φ indicates the operation of raising to the fifth power.

Functions of several variables are denoted similarly. For instance, let $z = x^2 - x2^y$. Here x and y are independent variables and z is regarded as a dependent variable. We can write $z = f(x, y)$ where the comma is a sign indicating the dependence of the function on two arguments. In this case particular values are found in the following way:

$$f(2, 1) = z \Big|_{\substack{x=2 \\ y=1}} = 2^2 - 2 \times 2^1 = 0;$$

$$f(1, 2) = z \Big|_{\substack{x=1 \\ y=2}} = 1^2 - 1 \times 2^2 = -3$$

and so on.

One must get used to this notation and operate on it quite freely. Here we give several examples of such operations. Suppose we have the functions $y = f(x) = x^2 - 3x$ and $z = \varphi(x) = 2x + 1$, and let a be a constant number. Then

$f(a) = a^2 - 3a$ (this is the value of the first function for $x = a$);

$\varphi(a^2) = 2a^2 + 1$ (this is the value of the second function for $x = a^2$);

$f(x^2) = (x^2)^2 - 3x^2 = x^4 - 3x^2$ [this is the value of y assumed when x^2 is substituted for the argument; we thus obtain a new function of x which may be denoted as $F(x)$];

$[f\,(x)]^2 = (x^2 - 3x)^2 = x^4 - 6x^3 + 9x^2$ (this is one more function of x);

$\varphi\,(x + a) = 2\,(x + a) + 1 = 2x + 2a + 1$ (this is another new function of x);

$f\,(x)\,\varphi\,(x) = (x^2 - 3x)\,(2x + 1) = 2x^3 - 5x^2 - 3x;$

$f\,(\varphi\,(x)) = [\varphi\,(x)]^2 - 3\varphi\,(x) = (2x + 1)^2 - 3\,(2x + 1) =$
$= 4x^2 - 2x - 2;$

$\varphi\,(f\,(x)) = 2f\,(x) + 1 = 2\,(x^2 - 3x) + 1 = 2x^2 - 6x + 1;$

$f\,(x + s) = (x + s)^2 - 3\,(x + s) = x^2 + 2xs + s^2 - 3x - 3s$

[this is a function of two variables which can be denoted by $\Phi\,(x, s)$] etc.

In particular, in the above examples we come across the operation of composing "a function of a function" or, as it is usually said, we deal with a **composite function.** A composite function is usually obtained in the following way. Let a variable y depend on a variable u and let u, in its turn, depend on a variable x. Thus, $y = f\,(u)$ and $u = \varphi\,(x)$. Then variations in x change u and therefore y also varies. Hence, y is a function of x of the form $y = f\,(\varphi\,(x))$. Thus we obtain a composite function. In this case u is an intermediate variable. There may also be several intermediate variables.

If we only want to designate that y is a function of x avoiding all the intermediate operations we can write $y = y\,(x)$. For instance, in the examples in Sec. 11 we could have written $S = S\,(R)$ or $p = p\,(V,\ T)$ and $V = V\,(T,\ p)$.

13. Methods of Representing Functions. If we intend to investigate a function, that is a dependence of one quantity upon another, the function must be represented in a certain way. There are several methods of representing functions.

The **analytical method** (i.e. representing a function by a formula) is one of the most widely used methods in mathematics. This method describes the mathematical operations which should be performed on the independent variable to obtain the value of the function. The operations are indicated by a formula. For example, the formula $y = x^2 - 2x$ says that in order to compute the value of the function y we must raise the corresponding value of the argument to the second power and then subtract the doubled value of the argument from the result.

The analytical method is compact (i.e. formulas usually occupy little space), it can be easily reproduced (i.e. it is not difficult to rewrite a formula). Besides, it is the most suitable method for performing mathematical operations on functions. Here we mean the algebraic operations (addition, multiplication and so on), the operations of higher mathematics (differentiation, integration and the like), and others. But the method is not visual enough (this means that when we have a formula it is not always possible to visualize the character of dependence of the function upon its argument).

The calculation of particular values of a function represented by a formula (in case the values are needed) may be a complicated operation. In addition, not all functions can be represented by a formula, and it may be inconvenient to put down a formula even when it exists.

It is sometimes necessary to use several different formulas to represent a function on different parts of the range of its argument. For example, let a material point fall without an initial velocity
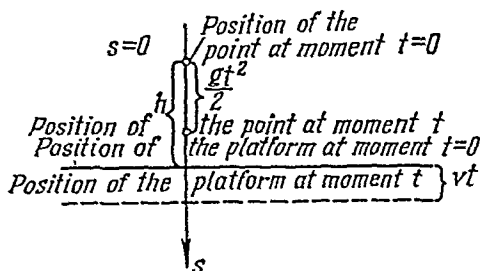
Fig. 6

on a platform which is moving downwards uniformly with the velocity v, the distance between the point and the platform at the moment $t = 0$ being equal to h. Then the path s covered by the point is a function of the time t, i.e. $s = f(t)$. According to Fig. 6 the relationship is determined by the formula

$$s = f(t) = \begin{cases} \dfrac{gt^2}{2} & (0 \leqslant t \leqslant t^*) \\ h + vt & (t^* \leqslant t < \infty) \end{cases}$$

where t^* is the moment of the impact of the point against the platform which can be found from the equation $\dfrac{gt^2}{2} = h + vt$.

The **tabular representation** of a function gives the numerical values of the function for certain discrete values of the argument. Such a table may have the following form:

$$y = f(x)$$

x	x_1	x_2	x_3	$\ldots$	x_N
y	$y_1 = f(x_1)$	$y_2 = f(x_2)$	$y_3 = f(x_3)$	$\ldots$	$y_N = f(x_N)$

(2)

Each of the differences $x_2 - x_1$, $x_3 - x_2$, $\ldots$ is called a **step of the** table. The tables with a **constant step** are the most convenient; an

argument x entering into such a table is taken for the values a, $a + h$, $a + 2h$, ... Here the constant step is denoted by h. The well-known tables of logarithms, of trigonometric functions etc. are examples of tabular representation of functions. In these tables, to save space, the values of the functions are written in lines (like words in a book) but not in a single row [as in (2)]. There are many other different tables of various important functions. Some tables represent a result of an experiment in which the values of one of the quantities are set beforehand and the values of the other quantities are measured etc.

The great advantage of the tabular method is that the values of a function are already calculated and therefore they can be immediately utilized. But we sometimes need the values of a function represented by a table which correspond to the values of the argument that are not included into the table. Then we have to perform some additional operations, namely, the operation of **interpolation** (i.e. calculating the values of the function for intermediate values of the argument) or the operation of **extrapolation** (i.e. calculating the values of the function for the values of the argument that fall outside the table). These additional calculations often yield incorrect results. Tables sometimes occupy much space. Building up a table usually requires much work. But in recent years the development of modern computer techniques has made it possible to calculate tables more quickly.

The disadvantage of the method is that it is inconvenient for performing mathematical operations because each new operation can make it necessary to compile a new table which is hard work. Besides, this cannot sometimes be done with a desired accuracy.

The third basic method of representing functions is the **graphical method**. The method represents a function by means of constructing its graph which enables us to visualize the character of variations of the function. Besides, when we have the graph of a function we can easily find approximate values of the function accurate to one or two decimal places but, of course, only in a given range of the argument. The construction of a more accurate graph requires much effort and yet the accuracy of the values of the function obtained by means of the graph may not be sufficient. It should be noted that some graphs characterizing an experiment may be drawn by a self-recording apparatus.

The fourth method of representing functions has been widely spread in recent years and is now becoming one of the most important methods. This is the **method of compiling a program** for calculating the values of a given function with the help of an electronic computer. We shall discuss the method in Sec. XIX.6.

All these methods in a certain sense supplement one another. We often come across the problem of passing from one method to another,

that is the problem of constructing a graph, of compiling a table (the so-called tabulation) or the problem of finding a suitable formula. Later on we shall discuss the problems of this kind.

There are, of course, many other ways of representing functions. We can sometimes give a verbal explanation of the law of correspondence between an independent variable and a function. For instance, we can say that a tax is such-and-such function of one's income.

For the first time the definition of the concept of a function close to the modern definition was given by the Swiss mathematician Johann Bernoulli (1667-1748). in 1718. But in the 18th century functions were usually understood in the sense of an analytical formula. The modern general concept of a function based on the notion of a law of interdependence between variables was first introduced by Euler in 1755 but it became universally recognized only in the 19th century.

14. Graphs of Functions. Graphs serve for the geometrical representation of functions. We shall remind the reader of the techniques of constructing graphs which are known from elementary mathematical courses. Let a variable y be a function of a variable x, i.e. $y = f(x)$. In order to construct the graph of the function we choose two number scales (axes) lying in a plane. The x-axis is usually drawn from left to right and is called the axis of abscissas and the y-axis is perpendicular to the x-axis and is called the axis of ordinates. The origin from which the coordinates are reckoned is often chosen at the intersection point of the axes (see Fig. 7). Then we make the argument take on different values and find the corresponding values of $y = f(x)$ which enables us to construct the points of the graph.

The point M shown in Fig. 7 is an arbitrary "moving" (variable) point of the graph and it has current coordinates x, y. Practically we cannot construct a very large number of points. We then connect the points with a curve and thus receive an approximate representation of the graph. But theoretically we interpret such a construction as if the variable x ran through its *whole* range; then the moving point M should run along the whole graph. Fig. 7 represents an example of a graph. We see that in this case the value of the function first increases as the argument x increases; such an increase lasts until x approaches, approximately, the value $x = 0.5$; then the function begins to decrease (comparatively slowly) and beginning with $x = 2$ the function increases again with an increasing rate.
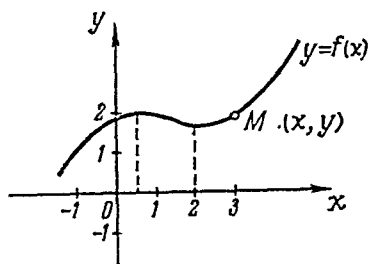
Fig. 7

The units of length and the origins on each of the axes should be chosen in such a way that all the most interesting peculiarities of the behaviour of the function on the corresponding intervals of the ranges of the argument and of the function should be represented most clearly.

For example, let us consider the graph of a function which describes a uniformly accelerated motion. Let the law of motion be

$$s = 98 + 0.01t^2 \quad (t \geqslant 0) \tag{3}$$

where t is measured in seconds and s is measured in centimetres. In this case we can choose the number scales on the axes in the way shown in Fig. 8. It is clear that a change of the position of the origin
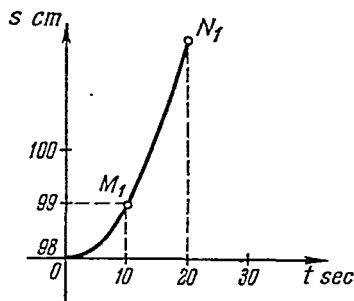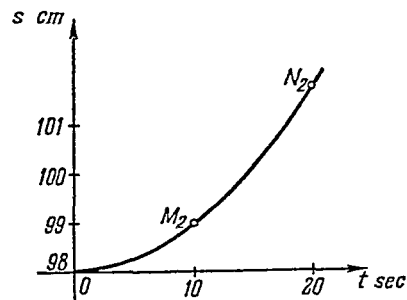
Fig. 8   Fig. 9

on the axis of the argument or on the axis of the function results, respectively, in the parallel translation of the graph as a whole along the x-axis or along the y-axis. If we increase or decrease the unit of length along one of the axes then the graph will, respectively, expand or contract along the same direction in such a way that the distances from all the points of the graph to the other axis will increase or decrease the same number of times. For instance, Fig. 9 represents the form of the graph of the same function (3) after the scale for the t-axis has been changed. The new graph is obtained from the original one by expansion in the direction parallel to the t-axis.

In order to represent the behaviour of a function in the best way one sometimes uses non-uniform scales which were already mentioned in Sec. 4.

In what follows we shall regard variables (both arguments and functions) as dimensionless unless the contrary is explicitly stated. In theoretical investigations the simplest thing is to take equal units of length for both axes and reckon the coordinates from the intersection point of the axes (which is then called the origin of the coordinate system) and we shall follow this way. We have already

described above in what way changes of scales or of the position of the origin affect the form of a graph.

15. **The Domain of Definition of a Function.** The domain of definition of a function is the totality of all the values of the independent variable for which the function is defined, that is the admissible range of the independent variable (see Sec. 5). We usually consider continuous variables and in such cases, as it was pointed out in Sec. 5, the domain of definition consists of one or several intervals.

The structure of the domain of definition of a function is sometimes implied by a physical or geometrical meaning of the function. For example, if we consider the relation $S = \pi R^2$ describing the dependence of the area of a circle upon its radius the domain of definition of the function is $0 < R < \infty$ since the geometrical meaning of R implies these very values. In case we consider the dependence of the atmosphere density ρ at a point (lying above a given point of the earth's surface at the height h above sea level) the domain of definition of the function $\rho = \rho\,(h)$ is the interval $h_0 \leqslant h \leqslant H$ where h_0 is the height of the earth's surface and H is a conditional height which is regarded as the limit of the atmosphere. If a function is represented by a formula then the set of all the values of the argument for which this formula gives a certain real value of the function is regarded as its domain of definition (as long as we consider *real functions of a real argument*, that is functions for which both the independent variable and the dependent one assume only real values). For instance, if $y = x^3$ then x can take on any real values, i.e. the domain of definition is the whole number line $-\infty < x < \infty$. If $y = \sqrt{x^2 - 2}$ then we cannot obtain real values of y while extracting the root in case we have $x^2 - 2 < 0$. Consequently, there must be $x^2 - 2 \geqslant 0$, i.e. $x^2 \geqslant 2$. The last inequality is fulfilled for $x \leqslant -\sqrt{2}$ or $x \geqslant \sqrt{2}$; the domain of definition consists of two intervals $-\infty < x \leqslant -\sqrt{2}$ and $\sqrt{2} \leqslant x < \infty$ (the domain is shaded in Fig. 10). In other analogous cases in order to determine the domain of definition of a function we must first find out what may prevent us from getting real values of the function and then form inequalities (as it was done in the last example when we put down $x^2 - 2 \geqslant 0$) which guarantee the possibility of obtaining real values. Then the problem of determining the domain of definition is reduced to solving these inequalities.

If an independent variable is discrete the domain of definition of the corresponding function consists of discrete (separate) points. For instance, if $f\,(x) = x! = 1 \times 2 \times \ldots \times x$ then x can assume only the values $1, 2, 3, \ldots$. In case a discrete argument takes on only integer values, as in the above example, it is usually denoted not by x but by the letters n, m, k and the like whereas the values $f\,(1)$, $f\,(2)$, $\ldots$, $f\,(n)$, $\ldots$ are denoted as a_1, a_2, $\ldots$, $a_n \ldots$.

In such a case we say that there is a **sequence**; for example, a geometric progression of the form

$$a_1 = a, \quad a_2 = aq, \quad a_3 = aq^2, \ldots, \quad a_n = aq^{n-1}, \ldots$$

is an example of a sequence etc. The graph of a function of a discrete argument is not a continuous line but consists of discrete points (see Fig. 11).

The range of a dependent variable, that is the set of all the values assumed by a function as its argument runs over the domain of definition of the function, is called the **range of the function**. For example, the domain of definition of the function $y = x^2$ is the interval $-\infty < x < \infty$ and the range of this function is the interval $0 \leqslant y < \infty$ since in this case y assumes only non-negative values.
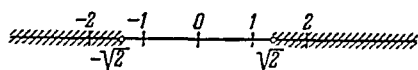
Fig. 10
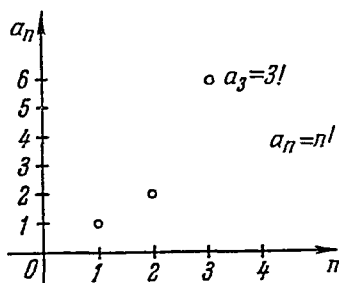
Fig. 11

The determination of the domain of definition of a function is essential for constructing its graph because this domain is just a part of the axis of abscissas over or under which the graph is placed. We see three simple graphs depicted in Fig. 12; the domains of the
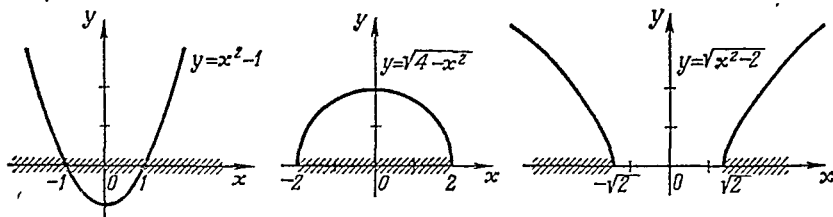
Fig. 12

functions are shaded. It is clear that in case a domain of definition consists of several separate parts the corresponding graph also consists of several components.

16. Characteristics of Behaviour of Functions. Now our aim is to study the ways of describing characteristic features of the behaviour of functions.

Unless otherwise stated, we shall regard functions in question as **single-valued**, that is we shall suppose that to each value of an

independent variable taken from the domain of definition of a function there corresponds only one certain value of the function. Multiple-valued functions will be discussed in Sec. 20.

A function is called **increasing (decreasing)** if the values of the function increase (decrease) as the argument increases. Both increasing and decreasing functions are called **monotonic**. When a function is not monotonic it is usually possible to indicate the **intervals of monotonicity** of the function on the axis of the argument; the function is monotonic on each of such intervals. Between the intervals of monotonicity there are often the **intervals of constancy** of the function. For example, in Fig. 13 we see the graphs of an increasing
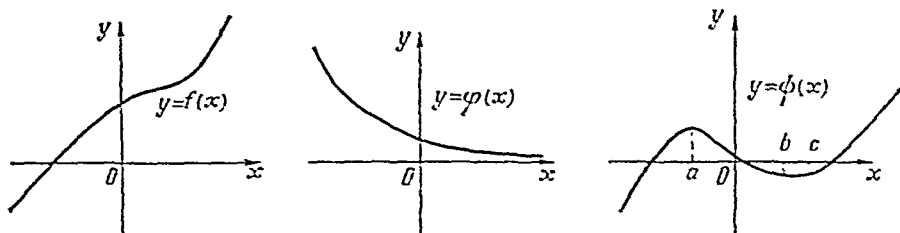
Fig. 13

function $f(x)$, of a decreasing function $\varphi(x)$ and of a non-monotonic function $\psi(x)$; the function $\psi(x)$ has the interval of increase $-\infty < x \leqslant a$, the interval of decrease $a \leqslant x \leqslant b$, the interval of constancy $b \leqslant x \leqslant c$ and the interval of increase $c \leqslant x < \infty$.

The condition that a function $f(x)$ increases can be written in the following way: $x_1 < x_2$ always implies $f(x_1) < f(x_2)$. This enables us to perform similar operations on both sides of inequalities involving the argument and the values of an increasing function: for example, as we know that $y = x^3$ is an increasing function we see that an inequality of the form $a < b$ implies $a^3 < b^3$ and vice versa. In case a function $f(x)$ is not monotonic the operation of this type can be performed on every interval of monotonicity in which the function increases. If a function decreases on some interval then $x_1 < x_2$ implies $f(x_1) > f(x_2)$. For instance, the function $y = x^2$ decreases over the interval $-\infty < x \leqslant 0$ and increases for $0 \leqslant x < \infty$; hence, $a < b$ implies $a^2 > b^2$ in case $b \leqslant 0$ and $a^2 < b^2$ if $a \geqslant 0$.

A function is called **continuous** if a continuous ("gradual") change of the argument results in a continuous change of the values of the function (without "jumps"). If otherwise the function is called **discontinuous** and every value of the argument for which the continuity ("gradualness") of the change of the function does not take place is called the **point of discontinuity** of the function. (These notions

will be discussed in detail in § III.4.) For example (see Fig. 14), the function $y = x^2$ is continuous over the whole x-axis; the function $y = \frac{1}{x}$ has one point of discontinuity $x = 0$ (the values of the function "approach infinity" as the values of the argument approach the
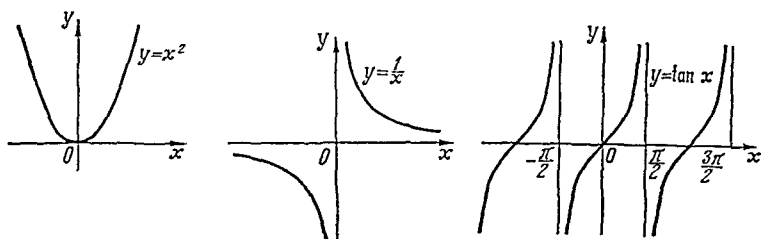
Fig. 14

value $x = 0$) and is continuous for all other values of $x \neq 0$; the function $y = \tan x$ has an infinitude of points of discontinuity

$$x = \pm \frac{1}{2} \pi, \pm \frac{3}{2} \pi, \pm \frac{5}{2} \pi, \ldots .$$

If a function is defined on both sides of its point of discontinuity the graph of the function is also discontinuous and consists of two or more pieces (components; see, for example, Fig. 14).
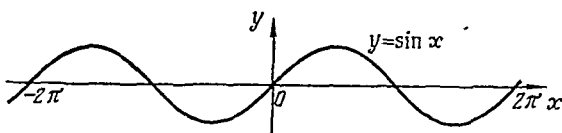
Fig. 15

The function $y = \sin x$ is an example of a periodic function. Namely (see Fig. 15), the behaviour of the function on the intervals

$$\ldots, \quad -4\pi \leqslant x \leqslant -2\pi, \quad -2\pi \leqslant x \leqslant 0,$$
$$0 \leqslant x \leqslant 2\pi, \quad 2\pi \leqslant x \leqslant 4\pi, \quad \ldots$$

is similar. More precisely, $\sin (x + 2\pi) \equiv \sin x$. Here the sign $\equiv$ is the **sign of identity**. We write it when we want to underline that an equality is an identity (it is also permissible to put down the usual sign of equality in such a case).

The number 2π is called the period of the function $y = \sin x$. We also have the identities

$$\sin (x + 4\pi) \equiv \sin x, \quad \sin (x + 6\pi) \equiv \sin x,$$
$$\sin [x + (-2\pi)] \equiv \sin x \quad \text{etc.}$$

But the period is usually understood as a positive number and even the least number for which the identities hold. Therefore it is 2π that is the period of $y = \sin x$ but not 4π, 6π etc. By the way, in order to indicate the fact we sometimes speak about the **primitive period**.

Generally, a function $y = f(x)$ is called a **periodic function** with period $A > 0$ if there is an identity of the form $f(x + A) \equiv f(x)$. Such a function behaves in the same way on each of the intervals

$$\ldots, \quad a - 2A \leqslant x \leqslant a - A, \quad a - A \leqslant x \leqslant a,$$
$$a \leqslant x \leqslant a + A, \quad a + A \leqslant x \leqslant a + 2A, \quad \ldots$$

where a is an arbitrary number. Therefore in order to investigate the function it is sufficient to consider its behaviour on one of the
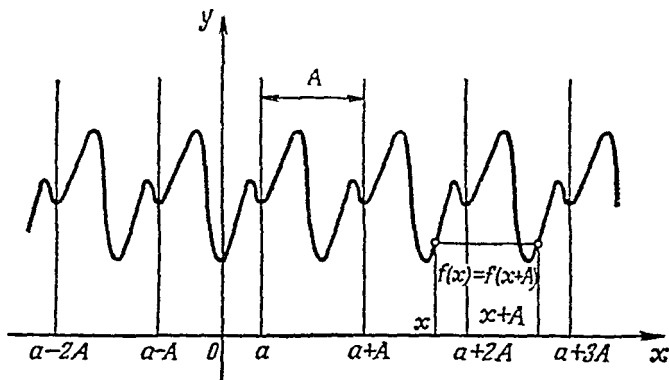
Fig. 16

intervals (see Fig. 16). The equality $f(x + A) = f(x)$ is illustrated in Fig. 16 for one of the values of x.

A function $y = f(x)$ is called an **even function** if it does not change its value when the sign of the argument is changed, that is if $f(-x) \equiv f(x)$. The examples of even functions are $y = x^2$, $y = x^6$, $y = \cos x$ etc. Fig. 17 shows that the graph of an even function is symmetric with respect to the axis of ordinates. A function $f(x)$ is called **odd** in case it is multiplied by -1 when the sign of the argument is changed, that is $f(-x) \equiv -f(x)$. The examples of odd functions are $y = x$, $y = x^5$, $y = \sin x$ etc. Fig. 18 illustrates the fact that the graph of an odd function is symmetric with respect to the origin of the coordinate system. It should be noted that in the general case a function may be neither even nor odd; for example, this is the case with the functions $y = 1 + \sin x$, $y = 1 - x$, $y = 2^x$. $y = \log x$ etc.

17. Algebraic Classification of Functions. Functions represented by a single formula (see Sec. 13) are classified depending on the

necessary algebraic operations which should be performed on the values of the argument in order to obtain the values of the functions. If only the operations of addition, subtraction and multiplication are used, and also the operation of raising to a positive integral power which is a special case of multiplication. the function is called a polynomial or an entire (integral) rational function; in
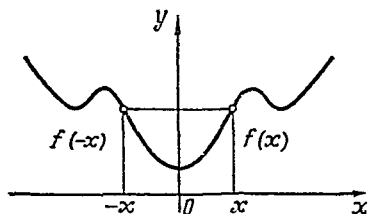
Fig. 17
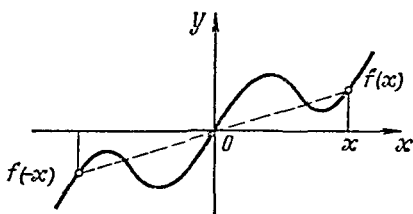
Fig. 18

forming a polynomial arbitrary constant coefficients can be used. Examples of polynomials are $y = x^3 - 2x + 3$, $y = x^2$, $y = 3$, $y = -\frac{x}{\pi} + \sqrt{2}x^3$, $y = a^4x^2 - 2$ and the like. On the other hand, the functions $y = x^{-5}$ and $y = x^3 + 2\sqrt{x}$ are not polynomials in the sense of the above definition. Every polynomial is characterized by its *degree* which is the highest of the exponents of powers of the independent variable entering into the expression of a polynomial; for instance, the degrees of the above written polynomials are, respectively, 3, 2, 0, 3, and 2.

Rational functions form a wider class of functions: these are the functions which involve the additional operation of division. If a rational function is not an entire function it is called a fractional rational function. An example of such a function is

$$y = \frac{x^2 - \dfrac{1}{x-1}}{x\sqrt{2}-3} - \frac{ax-\sqrt[3]{3}}{x^{-2}+1}$$

According to the rules of elementary algebra *every rational function can be represented as a ratio of two polynomials after all the summands entering into the expression of the function are reduced to a common denominator.*

There is a still wider class of functions whose analytical expressions may involve an additional operation of extracting roots. This is the class of **algebraic** functions. If an algebraic function is not rational it is called **irrational**. An example of an irrational function is $y = x^2 - \frac{1}{x} + \sqrt{x^2 - 1}$.

Functions which are not algebraic are called **transcendental**. Examples of transcendental functions are $y = \sin x$, $y = x^2 + $ $+ \tan x$, $y = 2^x$, $y = \log x$ etc. We point out that the last two functions are transcendental despite the fact that they are sometimes traditionally considered in elementary courses on algebra.

All these definitions are automatically extended to functions of several independent variables. The only new fact is the definition of the degree of a polynomial in several variables: it is defined as the greatest of the sums of the exponents of arguments entering into the monomials which are the summands in the expression of the polynomial.

For instance, the function $f(x, y) = x^4 y - x^4 y^2 + x$ is a polynomial of the sixth degree in x and y. But if we regard y as fixed the same function will be a polynomial of the fourth degree in x.

A polynomial of the first degree and a polynomial of the second degree are called, respectively, a **linear** function and a **quadratic** function. A polynomial of the third degree is called a **cubic** function and so on. These terms are applied for any number of independent variables.

18. Elementary Functions. We first enumerate the basic elementary functions studied in elementary mathematical courses:

$y = x^a$ (where a is constant) is a **power function**;

$y = a^x$ (where a is constant) is an **exponential function**;

$y = \log_a x$ (where a is constant) is a **logarithmic function**;

$y = \sin x$, $y = \cos x$, $y = \tan x$ and $y = \cot x$ are **trigonometric functions (circular functions)**;

$y = \arcsin x$, $y = \arccos x$ etc. are **inverse trigonometric functions (inverse circular functions)**.

Elementary functions are all the functions which can be obtained from basic elementary functions by means of algebraic operations (with any numerical coefficients) and the operation of composing a function of a function (see Sec. 12). In view of this definition all the algebraic functions are elementary. But very many transcendental functions are also elementary, for example, the functions $y = x + $ $+ \log \sin x$, $y = 2^{\log \tan x + \sin x}$ etc. (one may come across very complicated expressions of this type).

The class of elementary functions includes the greater part of functions treated in general courses of higher mathematics. As an example of a function which is not elementary we can mention $y = x!$ (but we do not give now the general definition of the function involving non-integral values of the argument. On this question see Sec. XIV.17). Many non-elementary functions are widely used in special branches of mathematics and its applications. Many of these functions have been investigated in detail and therefore the traditional classification into elementary and non-elementary functions may now be considered out of date.

19. Transforming Graphs. It often happens that we know the graph of a function and it is necessary to construct the graphs of some other functions which can be expressed in a certain way in terms of the former graph. Here we give several examples of transforming graphs in this manner.

Let the graph of a function $y = f(x)$ be given. It is required to construct the graphs of the functions $z = f(x) + a$ and $u = f(x+b)$ where a and b are some constants. The values of the quantities z and u will be represented on the same axis of ordinates as that for y (see Fig. 19). Then we have $z = y + a$ for any x and therefore
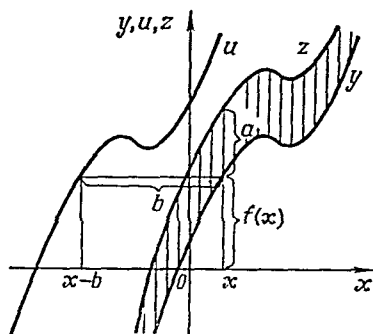
Fig. 19   Fig. 20
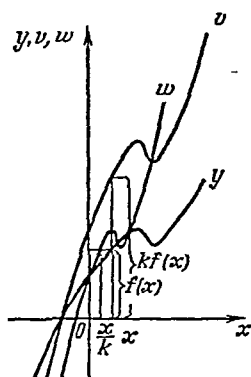
the graph of the function $z(x)$ can be obtained from the graph of the function $y(x)$ by translating the latter along the y-axis by the distance a in the positive direction of the axis in case $a > 0$. Such a translation is depicted in Fig. 19 where each of the vertical line segments has the length a. As for the graph of the function $u(x)$ one may think that it is obtained from the graph of $y(x)$ by translating the latter along the x-axis by the distance b in the positive direction in case $b > 0$. But this conclusion is wrong; in fact we should displace the graph of $y(x)$ by the amount b in the negative direction (if $b > 0$) in order to obtain the graph of $u(x)$. Indeed. the value $u = y$ is obtained if we take the value of the argument for u which is smaller by b than the corresponding value of the argument for y since $u = f[(x - b) + b] = f(x) = y$. Of course. if $a < 0$ or $b < 0$ the corresponding translation should be carried out in the opposite direction. On the other hand, when we say, for example. "to displace upwards by (-3)" we mean, in fact, "to displace downwards by $(+3)$" etc. Therefore it is permissible to say that we translate a graph in a certain direction by an amount h no matter what the sign of h is.

The graphs of the functions $v = kf\,(x)$ and $w = f\,(kx)$ are constructed in a similar way (see Fig. 20). The graph of the function $v\,(x)$ is obtained from the graph of $y\,(x)$ by the uniform k-fold expansion of the latter in the direction of the y-axis so that all the distances from the points of the graph of $y\,(x)$ to the x-axis should increase k times (in case $k > 1$). Indeed, the points of the graph of $v\,(x)$ which have the same abscissas as the points of the graph of $y\,(x)$ have the ordinates k times the ordinates of $y\,(x)$. The graph of the function $w\,(x)$ is obtained from the graph of the function $y\,(x)$ by the uniform contraction of the latter toward the y-axis with the k-fold decrease (in case $k > 1$) of all the distances of the points of the graph of $y\,(x)$ to the y-axis. Actually, we have $w\left(\dfrac{x}{k}\right) = f\left(k\,\dfrac{x}{k}\right) = f\,(x) =$ $= y\,(x)$. Of course, what we have said is literally true if $k > 1$. In case $0 < k < 1$ the expansion is replaced by the contraction and vice versa. But again, when we say, for example, "the $\left(\dfrac{1}{3}\right)$-fold expansion" we actually mean "the 3-fold contraction" and the like. Therefore we can say that we perform a k-fold expansion (or contraction) without specifying the magnitude of k. In conclusion we remark that if $k < 0$ we should additionally apply the operation of forming the corresponding mirror images of the graphs of $v\,(x)$ and $w\,(x)$ (for the function $v\,(x)$ the mirror image must be taken about the x-axis and for the function $w\,(x)$ the mirror image must be taken with respect to the y-axis).

Now combining the above results we can say that the graph of the function $y = kf\,(mx + b) + a$ can be obtained from the graph of the function $y = f\,(x)$ by means of the following transformations (performed in succession): the parallel translation along the x-axis [which yields the graph of $y = f\,(x + b)$], the contraction [which results in the graph of $y = f\,(mx + b)$], the expansion [which gives the graph of $y = kf\,(mx + b)$] and one more final translation along the y-axis resulting in the desired graph of $y = kf\,(mx + b) + a$. (If necessary, the corresponding mirror images should also be taken.)

The same results can be obtained by the corresponding operations on the coordinate axes without changing the graph. For example, instead of displacing the graph to the right we can translate the axes to the left, or, in other words, displace the origin (from which x is reckoned) to the left. Similarly, instead of expanding the graph and increasing the distances from the x-axis k times we can decrease the corresponding unit of length for the y-axis k times.

We can perform arithmetical operations on functions represented graphically. For example, Fig. 21 illustrates the graphical addition of two functions: the graphs of $f\,(x)$ and $\varphi\,(x)$ are given and it is necessary to construct the graph of their sum. The equal line segments lying one above another are shown in heavy lines.

of the axes x and y interchanged (see Fig. 23). Each of the single-valued branches is represented by the corresponding half of the parabola: the first branch is represented by the upper half and the second one by the lower half.

The graph of an implicit function may have the form shown in Fig. 24. Here we see that the function is single-valued for $x < a$ and for $x > b$ whereas it is three-valued for $a < x < b$. When separating the function into branches it is natural to regard the arc AB as the graph of the first branch, the arc BC as the graph of the second branch and the arc CD as that of the third one.
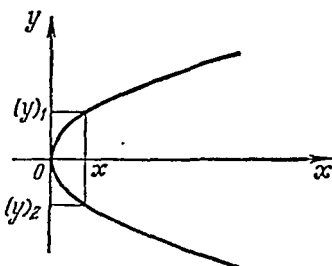
Fig. 23
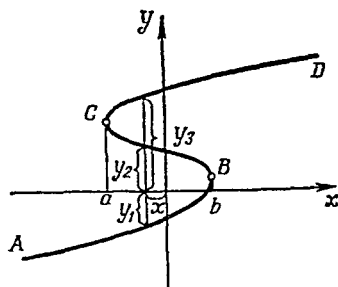
Fig. 24

In connection with the question of multiple-valued functions discussed above we can note that for some functions to every value of the independent variable there corresponds an entire interval of values of the function. For instance, the relation between the height of a person and his possible weight is an example of such a functional relation. Functions of this kind are usually investigated in the theory of probabilities (see Sec. XVIII.16) and they will not occur in other chapters of our course.

21. Inverse Functions. Suppose we are given a function

$$y = f(x) \qquad (6)$$

now take different values of y and find the corresponding that is let us choose the former dependent variable as and regard the former independent variable as a func-functional relation) $x(y)$ thus obtained is called of the original function $y(x)$. It is represented and (6) in which y is now regarded as an independent is regarded as a dependent variable. But we out that it is permissible to use different variables in considering one and the same

20. Imp. is defined b, the function. same function, $x - y^3 + 2 = $ same function $y = $ function implicitly represents the functi

function. Therefore, if we wanted to denote, as we usually did, the independent variable by x and the dependent variable by y for the inverse function we should simply have to substitute x for y and y for x in (6). Hence, using the new notation we rewrite the relation which defines the inverse function in the form

$$x = f(y) \qquad . \qquad (7)$$

Thus, the inverse function turns out to be represented in an implicit form and therefore (see Sec. 20) it is, generally speaking, multiple-valued. We can easily establish a condition which guarantees the
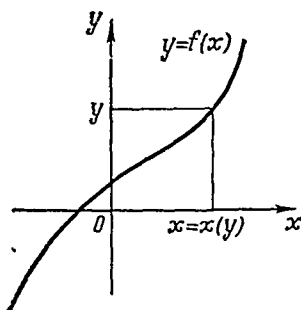
Fig. 25
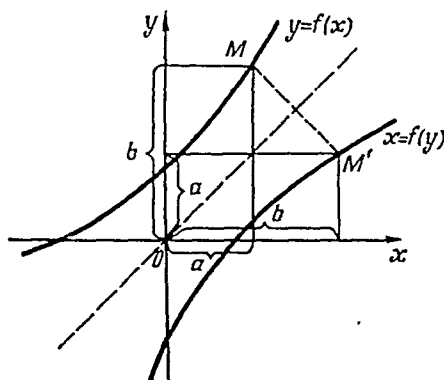
Fig. 26

single-valuedness of an inverse function: this is the monotonicity of the original function. Indeed, this being so, we obtain a certain uniquely defined value $x = x(y)$ for each given value of y (see Fig. 25).

Examples. The inverse function of the function $y = x^3$ is defined by the equality $x = y^3$, that is $y = \sqrt[3]{x}$; the inverse function of $y = x^2$ is the two-valued function $y = \pm \sqrt{x}$.

Equalities (6) and (7) differ only in the interchange of the notation of the quantities x and y, that is in the interchange of their roles. Therefore we see, as it is shown in Fig. 26, that the graph of the inverse function is obtained as a mirror image of the graph of the original function about the bisector of the angle between the coordinate axes (the bisector is represented by the dotted line in Fig. 26). The points M and M' in Fig. 26 both correspond to one and the same equality of the form $b = f(a)$.

We remark in conclusion that if the function $x(y)$ is the inverse function of the function $y(x)$ then, conversely, the latter is the inverse function of the former.

§ 4. Review of Basic Functions

Many of the functions which we are going to discuss here are studied in elementary mathematical courses. We shall consider them here because of their significance.

22. Linear Function. The general form of a linear function (see the end of Sec. 17) is

$$y = ax + b \qquad (8)$$

where a and b are constant coefficients.

The graph of a linear function is a straight line (see Fig. 27). The coefficient a is called the **slope** of the straight line. The greater $|a|$ (i.e. the greater a in its absolute value), the steeper the slope of the straight line (with regard to the x-axis). If the argument of a function changes from a value x_0 to a certain value x it receives an increment Δx (which is equal to $x - x_0$)*. Then the function receives the corresponding increment Δy [which is equal to $y - y_0 = = f(x) - f(x_0)$]. In our case $y = f(x) = ax + b$ and therefore $y_0 = ax_0 + b$ and $y = ax + b$. Consequently, $y - y_0 = a(x - x_0)$, i.e. $\Delta y = a \Delta x$. This implies

Fig. 27

$$\frac{\Delta y}{\Delta x} = a \quad \text{(if } \Delta x \neq 0) \qquad (9)$$
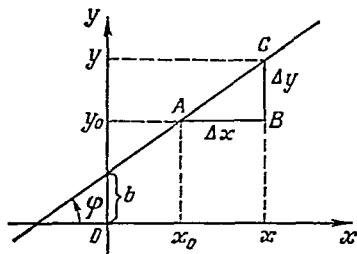
Thus, the ratio of the increment of a linear function to the increment of the argument is constant and equal to the slope of the graph. *The increment of a linear function is directly proportional to the increment of the argument.*

In Fig. 27 the case when $a > 0$ is shown. If $a < 0$ the straight line is drawn downwards to the right (see Fig. 28). In case $a = 0$ the straight line is parallel to the x-axis; in this case the function is constant and thus we obtain the graph of a constant.

The property of the increment of a linear function forms the basis for the so-called **linear interpolation** which is used even in elementary mathematical courses. The idea of this method is the following. Suppose we know the values of a function $y = f(x)$ (its graph is depicted in Fig. 29 by the dotted line) for $x = x_0$ and for $x = x_0 + h$:

$$f(x_0) = y_0, \quad f(x_0 + h) = y_1$$

---

* The Greek letter Δ (delta) is used to denote an increment. The symbol Δx should be regarded as an indivisible symbol and by no means as the product of Δ by x. An "increment" is understood in the algebraic sense, i.e. it can be positive, negative or equal to zero.

but the intermediate values of the function for the values of x lying between $x = x_0$ and $x = x_0 + h$ are unknown. Then we approximately replace the given function by a linear function which assumes the same values for $x = x_0$ and for $x = x_0 + h$, that is we replace the arc $\cup AE$ by the straight line segment AE. The similarity of the triangles ABC and ADE then implies

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{h}$$

i.e.

$$y = y_0 + \frac{y_1 - y_0}{h} (x - x_0)$$

Such a replacement is possible in case the function $f(x)$ slightly differs from the linear function on the interval between x_0 and $x_0 + h$. The interpolation method is widely used, in particular,
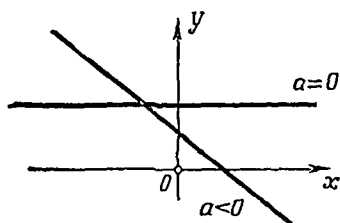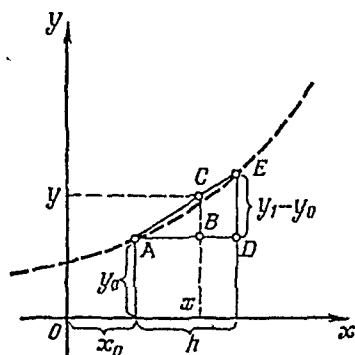
Fig. 28   Fig. 29

for tables with a sufficiently small step when the successive values of the function differ slightly from each other. More precise methods of interpolation will be discussed in Secs. V.6-8. The **linear extrapolation** (see Sec. 13) is performed in like manner.

Formula (9) and Fig. 27 imply that $a = \tan \varphi$, i.e. *the slope of a straight line is equal to the tangent of the angle of inclination of the line to the axis of abscissas.*

If the quantities x and y have certain dimensions the slope also has a dimension. Formula (8) shows that $[b] = [y]$ and $[ax] = [y]$ which implies $[a] = \frac{[y]}{[x]}$ (the dimensions of coefficients entering into other formulas can be determined similarly). The geometrical meaning of the slope can be easily interpreted in the general case: if l_x units of length for the x-axis correspond to the unit measure of the quantity x and l_y units of length for the y-axis correspond to

the unit measure of the quantity y (l_x and l_y are the so-called scale factors) then the sides AB and BC of the triangle ABC in Fig. 27 are of lengths $l_x \Delta x$ and $l_y \Delta y$, respectively. Consequently,

$$\tan \varphi = \frac{l_y \Delta y}{l_x \Delta x} \quad \text{and} \quad a = \frac{\Delta y}{\Delta x} = \frac{l_x}{l_y} \tan \varphi \qquad (10)$$

i.e. the slope is proportional to the tangent of the angle φ.

23. Quadratic Function. The general form of a quadratic function is

$$y = ax^2 + bx + c$$

From elementary mathematical courses it is known that *the graph of a quadratic function is a parabola*. In the simplest case when $a = 1$, $b = 0$ and $c = 0$, i.e. $y = x^2$, the graph has the form depicted in Fig. 30. Then the function is even and the y-axis is the symmetry axis of the graph (the axis of the parabola). The intersection point
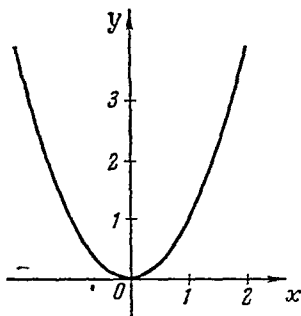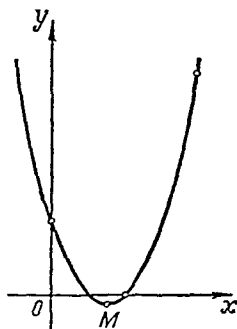
Fig. 30   Fig. 31

of a parabola with its axis is called the vertex of the parabola. The vertex in Fig. 30 is placed at the origin of the coordinate system.

In the general case when $a \neq 0$, b and c are arbitrary numbers the parabola is obtained from the parabola depicted in Fig. 30 by the operations of uniform expansion and parallel translation. To determine the position of the vertex we can apply the so-called method of completing a square which we shall demonstrate here by considering a concrete numerical example. Let the quadratic function $y = 2x^2 - 3x + 1$ be given*. Then we perform the following

---

* In practice we usually have quadratic functions (trinomials) whose coefficients are approximate numbers (in contrast to the above trinomial with exact coefficients). For instance, we can take the trinomial $y = 2.17x^2 - 3.21x + 0.84$ and the like. But if we investigate the case with exact coefficients we can easily pass to a more complicated case. The comment also refers to further examples of this type.

simple transformations:

$$y = 2\left(x^2 - \frac{3}{2}x + \frac{1}{2}\right) = 2\left[\left(x - \frac{3}{4}\right)^2 - \left(\frac{3}{4}\right)^2 + \frac{1}{2}\right] = 2\left(x - \frac{3}{4}\right)^2 - \frac{1}{8}$$

Consequently (see Sec. 19) we obtain the sought-for graph from the parabola depicted in Fig. 30 by translating the parabola $\frac{3}{4}$ unit of length to the right, expanding it along the direction of the y-axis with the two-fold increase of the distances from the points of the graph to the x-axis and, finally, by translating $\frac{1}{8}$ unit downwards. The graph thus obtained is
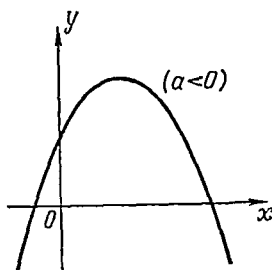
Fig. 32
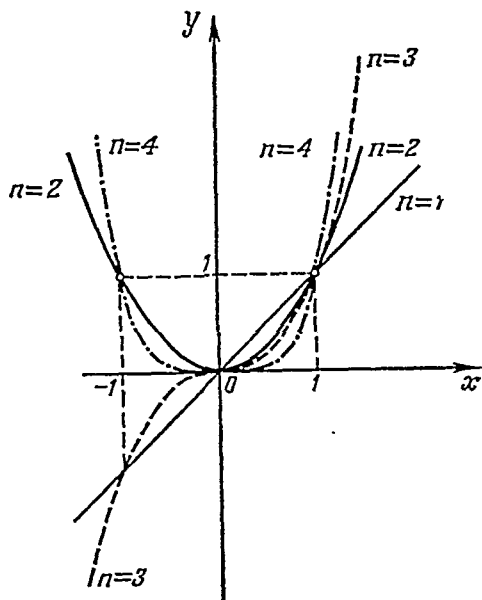
Fig. 33

shown in Fig. 31. To construct a more accurate graph we can additionally take several values of x and determine the corresponding values of y which enables us to construct the corresponding points of the graph. For instance, we have $y = 1$ for $x = 0$. $y = 0$ for $x = 1$ and $y = 3$ for $x = 2$; the corresponding points are indicated on the graph. The vertex of the constructed parabola is situated at the point M with the coordinates $x = \frac{3}{4}$ and $y = -\frac{1}{8}$. The parabola is "narrower" than the one depicted in Fig. 30 (with the same unit of length).

Generally, the greater $|a|$, the narrower the parabola. If $a < 0$ the branches of the parabola are open downwards (see Fig. 32). In case $a = 0$ the quadratic function turns into a linear one.

24. **Power Function.** The general form of a power function is

$$y = x^n$$

If $0 < x < 1$ then the greater n, the smaller the values of the function. But if $x > 1$ then the greater n, the greater the values of the function. In Fig. 33 we see the graphs for $n = 1, 2, 3$ and 4. While constructing the parts of the graphs of $y = x^n$ for the values $x < 0$ one should take into account that the function $y = x^n$ is even for even n and odd for odd n. In particular, let us consider in detail the graph of the function $y = x^3$ (the cubic parabola). The graph is convex upwards (or concave downwards) for $x < 0$, that is it lies under the tangent drawn at any of its points. For $x > 0$ the graph
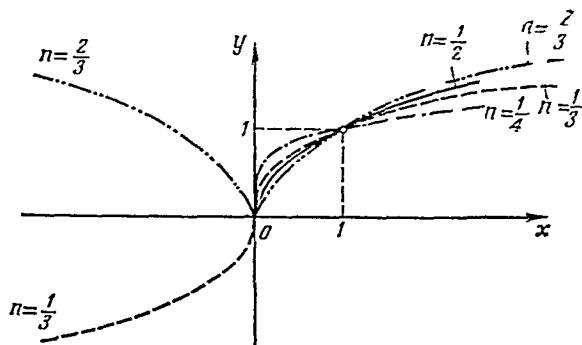
Fig. 34

is convex downwards (or concave upwards). If we pass from left to right through the origin of the coordinate system the direction of convexity changes to the opposite one. The tangent to the graph at the origin coincides with the x-axis but at the point of tangency O the curve passes from one side of the tangent line to another. Such points are called the points of inflection of a curve. Thus, the cubic parabola has one point of inflection. Among the well-known curves, the sinusoid, for example, has points of inflection. For fractional values of the exponent n the graphs are placed between the corresponding graphs for integral values of n. But in the case of a fractional n one should be careful when constructing graphs: a negative number raised to a fractional power may result in an imaginary number and therefore in such a case we must not construct the graph for $x < 0$.

Let us consider the case $0 < n < 1$. For instance, let $n = \frac{1}{2}$, that is $y = x^{\frac{1}{2}} = \sqrt{x}$. Then, as it was shown in Sec. 20, the graph is the upper half of the ordinary (quadratic) parabola with the symmetry axis coinciding with the x-axis (see Fig. 34). The graphs of power functions for some other fractional n are also shown in Fig. 34.

In case the fraction representing n has an odd denominator the graph exists not only for $x > 0$ but for $x < 0$ as well because, for negative numbers, we can extract real roots with odd indices of radicals. In particular, let us take the graph of the function $y = x^{\frac{2}{3}}$ (the semi-cubical parabola) depicted in Fig. 35. The graph first approaches
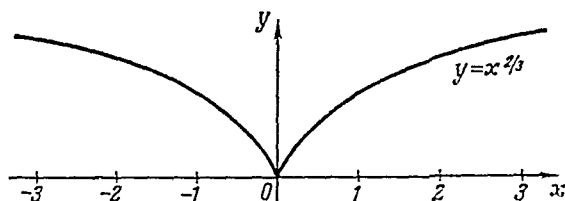
Fig. 35

the origin of the coordinates (for example, when we pass from left to right) and then departs from the origin. At the origin this curve has the so-called **spinode**, or **cusp**. Later on we shall investigate some other curves having cusps.

Finally, let us take the case of a negative n ($n = -m < 0$). Then $y = \dfrac{1}{x^m}$ and therefore we have very large values of $|y|$ for very small $|x|$ and vice versa. The corres-
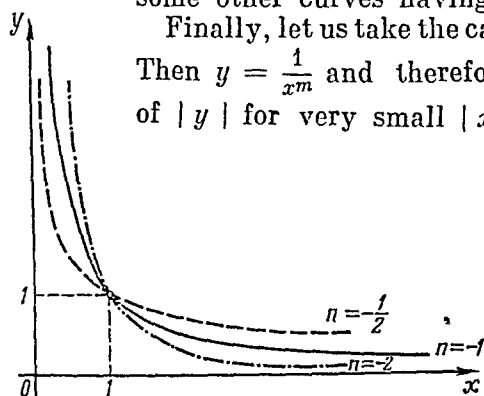
Fig. 36
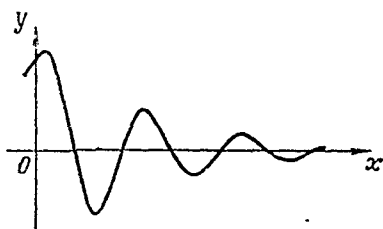
Fig. 37

ponding graphs are shown in Fig. 36 for $x > 0$; we leave to the reader the construction of the parts of the graphs corresponding to $x < 0$. All these graphs stretch along the coordinate axes and approach them unlimitedly as x or y approaches infinity. Generally, when a curve and a straight line have a mutual disposition of such a kind the straight line is called the **asymptote** of the curve. Hence, each of the above graphs has two asymptotes which are the coordinate axes.

One must not think that a curve cannot intersect its asymptote in all other cases. For example, investigating the so-called damped oscillations we obtain a graph of the form shown in Fig. 37. Here the x-axis is the asymptote of the graph which intersects it infinitely many times.

25. Linear-Fractional Function. A linear-fractional function has the general form

$$y = \frac{ax + b}{cx + d} \tag{11}$$

In the simplest case when $a = d = 0$ we obtain, denoting $\frac{b}{c} = k$, the expression $y = \frac{k}{x}$ which describes the inverse proportional relation. The corresponding graph, as is well known from elementary
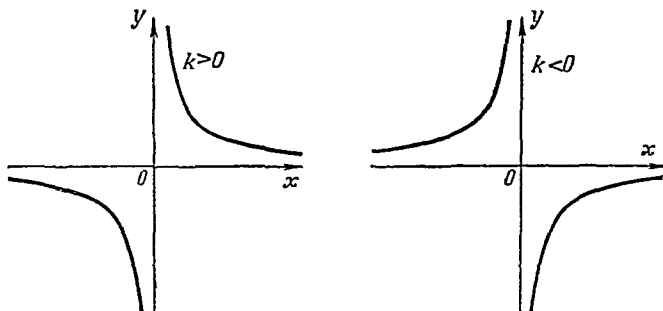
Fig. 38

mathematical courses, is called a **hyperbola**. The graph is depicted in Fig. 38 for the two cases $k > 0$ and $k < 0$ separately. The function $y = \frac{k}{x}$ being odd, the hyperbola has a centre of symmetry (which is located at the origin of coordinates in Fig. 38). It has two asymptotes (which are the coordinate axes in Fig. 38). We shall prove in Sec. II.13 that a hyperbola has two symmetry axes (for the hyperbola in question the axes of symmetry are the bisectors of the angles between the coordinate axes in Fig. 38).

In the general case the graph of a linear-fractional function is also a hyperbola which can be obtained by a parallel translation of the hyperbola depicted in Fig. 38. We shall illustrate this by taking a concrete numerical example. Let $y = \frac{2x + 3}{3x - 5}$. We now

carry out the following simple transformations:

$$y = \frac{2\left(x+\dfrac{3}{2}\right)}{3\left(x-\dfrac{5}{3}\right)} = \frac{2}{3}\,\frac{x-\dfrac{5}{3}+\dfrac{5}{3}+\dfrac{3}{2}}{x-\dfrac{5}{3}} = \frac{2}{3}\,\frac{\left(x-\dfrac{5}{3}\right)+\dfrac{19}{6}}{x-\dfrac{5}{3}} =$$

$$= \frac{2}{3}\left(1+\frac{\dfrac{19}{6}}{x-\dfrac{5}{3}}\right) = \frac{\dfrac{19}{9}}{x-\dfrac{5}{3}}+\frac{2}{3}$$

Thus we conclude that the desired graph can be obtained from the graph of the function $y = \dfrac{\dfrac{19}{9}}{x}$ by translating the latter $\dfrac{5}{3}$ units
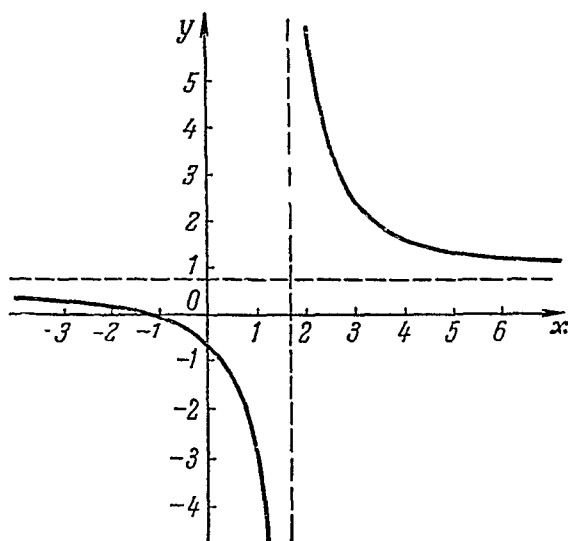
Fig. 39

of length to the right and $\dfrac{2}{3}$ units upwards. Hence, we get a hyperbola whose centre of symmetry is at the point $x = \dfrac{5}{3}$, $y = \dfrac{2}{3}$ (see Fig. 39).

A linear-fractional function of general form (11) has a point of discontinuity at $x = -\dfrac{d}{c}$ because the denominator vanishes at the point. This accounts for the fact that the graph consists of two separate portions (see Sec. 16).

*

26. Logarithmic Function. A logarithmic function is a function of the form

$$y = \log_a x \tag{12}$$

It is defined only for $x > 0$, and we consider the bases of logarithms $a > 0$ $(a \neq 1)$. The graphs of logarithmic functions are shown for different bases in Fig. 40. They have neither symmetry axes nor centres of symmetry but have an asymptote which is the y-axis. All the logarithmic functions are proportional to each other since taking logarithms to the base b of both sides of the equality $a^{\log_a x} = x$ we get

$$\log_b x = \log_a x \cdot \log_b a = k \log_a x \quad \left(k = \log_b a = \frac{1}{\log_a b} \right) \tag{13}$$

Therefore we can obtain all the graphs depicted in Fig. 40 by expanding or contracting one of them along the direction of the

Fig. 40

y-axis with a uniform increase or, respectively, with a uniform decrease of the distances of the points of the graph from the x-axis. Now let us consider the angles of intersection of the graphs with the x-axis. According to the general definition of the angle between intersecting curves as the angle between the tangents to the curves at the point of their intersection, we mean here the angles formed by the tangents to the graphs with the x-axis. When the graph is expanded or contracted in the way described above the tangent rotates about the point of intersection. We see that the tangent has a very slant inclination (to the x-axis) for very large values of a and a very steep inclination for values of a close to 1. For a certain value of a the angle of intersection of the graph of logarithmic function (12) with the x-axis is equal to 45°. *This value of a is denoted by the letter e.* It plays an important role in mathematics as we shall see later.

We see in Fig. 40 that the angle of intersection is greater than 45° for $a = 2$ and smaller than 45° for $a = 4$; hence, the number e lies between the limits 2 and 4. More accurate calculations which will be described in Sec. IV.16 show that $e = 2.71828$ with an accuracy of 10^{-5}. The notation e for this number was introduced by Euler.

Logarithms to the base e are called **natural logarithms** [Napierian **logarithms** after the Scottish mathematician J. Napier (1550-1617)].

They are denoted as $\ln x = \log_e x$. The graph of the natural logarithm is shown in Fig. 41. A logarithm to any other base can be expressed in terms of the natural logarithms in accordance with formula (13):

$$\log_a x = \frac{\ln x}{\ln a} \qquad (14)$$

Hence, the formulas for passing from the common (decimal) logarithms to the natural ones and vice versa are

$$\log x = 0.4343 \ln x \text{ and } \ln x = 2.303 \log x$$

where the values of the proportionality factors are accurate to four decimal places.

Besides the natural logarithms we also use the common logarithms (in numerical calculations) and the logarithms to the base 2 (in information theory and some other branches of modern mathematics).
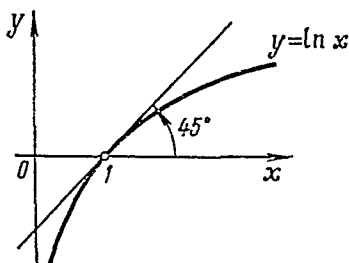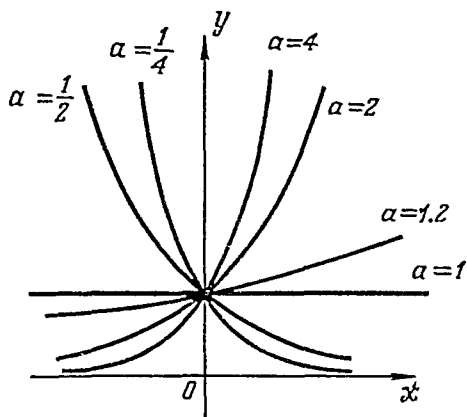
Fig. 41   Fig. 42

27. Exponential Function. An exponential function is a function of the form

$$y = a^x \qquad (15)$$

The function is defined for all x, and we always consider the values $a > 0$ (because raising $a < 0$ to a fractional power may result in an imaginary number). Equality (15) can be obtained from formula (12) if we solve it for x (which yields $x = a^y$) and then interchange x and y. Consequently (see Sec. 21) the exponential function and the logarithmic function are the inverse functions with respect to each other.

Therefore the graphs of exponential functions which are depicted in Fig. 42 for different bases a are obtained as the mirror images of

the corresponding graphs (shown in Fig. 40) with respect to the bisector of the angle between the coordinate axes. If $a > 1$ the exponential function is an increasing function, and the greater a, the greater the rate of its increase. In case $0 < a < 1$ the exponential function is a decreasing function.

We often deal with the exponential function with $a = e$. In this case there is special notation: $y = e^x = \exp x$.

Any exponential function with an arbitrary base a can be reduced to the base e; indeed, the definition of a logarithm implies that $a = e^{\ln a}$ and therefore $a^x = (e^{\ln a})^x = e^{kx}$ where $k = \ln a$.

28. Hyperbolic Functions. The hyperbolic sine, cosine and tangent are, respectively, the functions

$$\sinh x = \frac{e^x - e^{-x}}{2}, \quad \cosh x = \frac{e^x + e^{-x}}{2}, \quad \tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

At first these terms sound strange but their genuine sense (for example, the connection between $\sin x$ and $\sinh x$ or between the hyperbolic functions and a hyperbola) will be explained only when we get to Secs. VIII.4 and XIV.8. Let us now establish some formulas connecting these functions. Squaring the first two equalities we get

$$\sinh^2 x = \frac{e^{2x} - 2 + e^{-2x}}{4} \quad \text{and} \quad \cosh^2 x = \frac{e^{2x} + 2 + e^{-2x}}{4}$$

Now subtracting and adding these two formulas we obtain

$$\cosh^2 x - \sinh^2 x = 1, \quad \cosh^2 x + \sinh^2 x = \frac{e^{2x} + e^{-2x}}{2} = \cosh 2x$$

The obtained formulas indicate a significant analogy between hyperbolic and trigonometric functions. We leave to the reader the deduction of the formulas

$$\sinh 2x = 2 \sinh x \cosh x,$$
$$\sinh (a + b) = \sinh a \cosh b + \cosh a \sinh b,$$
$$1 - \tanh^2 x = \frac{1}{\cosh^2 x}$$

and other similar formulas. Note that $\sinh 0 = 0$ and $\cosh 0 = 1$. The functions $\sinh x$ and $\tanh x$ are odd whereas the function $\cosh x$ is even; indeed, for example,

$$\sinh (-x) = \frac{e^{(-x)} - e^{-(-x)}}{2} = \frac{e^{-x} - e^x}{2} = -\frac{e^x - e^{-x}}{2} = -\sinh x$$

The construction of the graphs of $\sinh x$ and $\cosh x$ is illustrated in Fig. 43. The graph of $\tanh x$ is shown in Fig. 44. To construct this graph we find its points with the help of the graphs of $\sinh x$ and $\cosh x$. It is clear that the hyperbolic functions do not possess the most important property of trigonometric functions, namely,

the periodicity. Besides, the range (see Sec. 15) of each hyperbolic function considerably differs from the range of the corresponding trigonometric function.
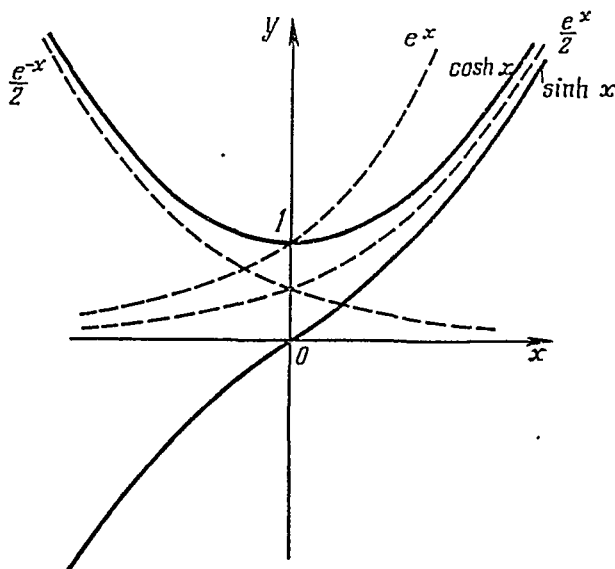
Fig. 43

The graph of $\tanh x$ has two asymptotes since, for large values of $|x|$, we have $e^{-|x|} \ll 1 \ll e^{|x|}$ (the sign $\ll$ means here "much smaller") and therefore $\tanh x \approx 1$ for large $|x|$ and $x > 0$, and $\tanh x \approx -1$ for large $|x|$ and $x < 0$.
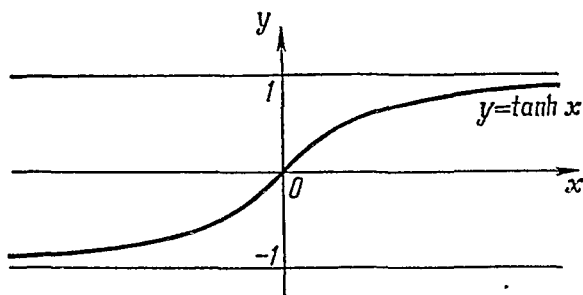
Fig. 44

We sometimes consider the **inverse hyperbolic functions** which are denoted as $\sinh^{-1} x$, $\cosh^{-1} x$ and $\tanh^{-1} x$, respectively. Figs. 43 and 44 show that the first and the third functions are single-valued (compare with Fig. 25) whereas the second one is two-valued. All

these functions can be expressed in terms of logarithms. In fact, let;
for example, $y = \sinh^{-1} x$. Then, by the definition of an inverse
function, we have

$$x = \sinh y = \frac{e^y - e^{-y}}{2}$$

i.e.

$$e^y - e^{-y} - 2x = 0, \quad e^{2y} - 2xe^y - 1 = 0$$

which implies $e^y = x \pm \sqrt{x^2 + 1}$. The left-hand side being positive,
the right-hand side should also be positive. Therefore we can take
only "$+$" in front of the radical. Now taking logarithms we obtain

$$y = \sinh^{-1} x \equiv \ln (x + \sqrt{x^2 + 1}) \tag{16}$$

29. **Trigonometric Functions.** The function $y = \sin x$ with
period 2π is well known from courses on trigonometry. Its graph
(the sinusoid, the sine curve) is represented in Fig. 45. The function
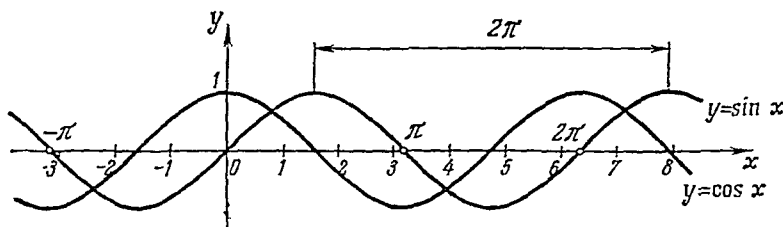
Fig. 45

is odd, has no points of discontinuity and is bounded (its values
lie between the limits -1 and $+1$). We have $\cos x = \sin \left(x + \frac{\pi}{2}\right)$
and therefore the graph of the function $\cos x$ is the same sinusoid
but translated $\frac{\pi}{2}$ units of length to the left; this graph is also shown
in Fig. 45. In many applications we encounter a sinusoidal, "harmonic" relation of the form

$$y = M \sin (\omega t + \alpha) \tag{17}$$

where the independent variable t is interpreted as time, the constant $M > 0$ is called an **amplitude** and $\omega > 0$ is called a frequency
(circular, angular frequency). The sum $\omega t + \alpha$ is called a phase
and the constant α is an **initial phase** which is obtained from the
phase by substituting $t = 0$ for t. We can easily investigate in what
way parameters M, ω and α affect the form and the disposition
of the sinusoid (compare with Sec. 19). The amplitude M increases
the range of the sinusoid and brings it from $-M$ to M, the frequency

ω changes the period 2π into $T = \dfrac{2\pi}{\omega}$ and the presence of the initial

phase α displaces the sinusoid to the left by the distance $\dfrac{\alpha}{\omega}$ [since

$\omega t + \alpha = \omega \left(t + \dfrac{\alpha}{\omega} \right)$ and therefore the value $\dfrac{\alpha}{\omega}$ is added to

the argument]. The graph thus obtained is represented in Fig. 46.

A function of form (17) is obtained, in particular, when we transform the expression $A \cos \omega t + B \sin \omega t$. The right-hand side of (17) can be rewritten in the form $M \sin \alpha \cdot \cos \omega t + M \cos \alpha \cdot \sin \omega t$ and thus in order to obtain the equality

$$A \cos \omega t + B \sin \omega t \equiv$$
$$\equiv M \sin (\omega t + \alpha) \quad (18)$$

we must have $A = M \sin \alpha$ and $B = M \cos \alpha$. From this it is easy to find M and α: $M = \sqrt{A^2 + B^2}$ and $\tan \alpha = \dfrac{A}{B}$; the quarter in which α should be taken is
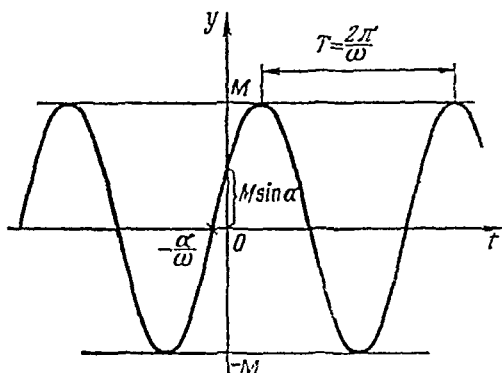
Fig. 46

defined by the signs of $\sin \alpha$ and $\cos \alpha$, i.e. by the signs of A and B.

In case the independent variable is interpreted not as time but as a geometrical coordinate the sinusoidal relation is usually written in the form $y = M \sin (kx + \alpha)$ instead of (17). In this case k is called a **wave-number** and $\lambda = \dfrac{2\pi}{k}$ is a **wave-length**.

The function $y = \tan x$ has the period π since $\tan (x + \pi) \equiv \tan x$. It has the points of discontinuity at $x = \dfrac{\pi}{2}, \dfrac{\pi}{2} + \pi$, $\dfrac{\pi}{2} - \pi$, ... (this can be written in the general form as $x = \dfrac{\pi}{2} + k\pi$ where $k = 0, \pm 1, \pm 2, \ldots$). Indeed, at these points $\cos x = 0$ and therefore $\tan x = \pm \infty$. The graph of the function (the tangent curve) is represented in Fig. 47; it consists of an infinitude of similar components and has infinitely many asymptotes. The graph of the function $y = \cot x$ is also shown in Fig. 47. We have

$$\cot x = -\tan \left(x - \dfrac{\pi}{2} \right)$$

and therefore the curve has the same form but its disposition is changed in the corresponding way.

The function $y = \text{Arc} \sin x$ is the inverse function of $y = \sin x$ and therefore its graph (see Fig. 48) is the mirror image of the graph of $y = \sin x$ about the bisector of the first quadrant angle. This function is multiple-valued (more precisely, infinite-valued) and
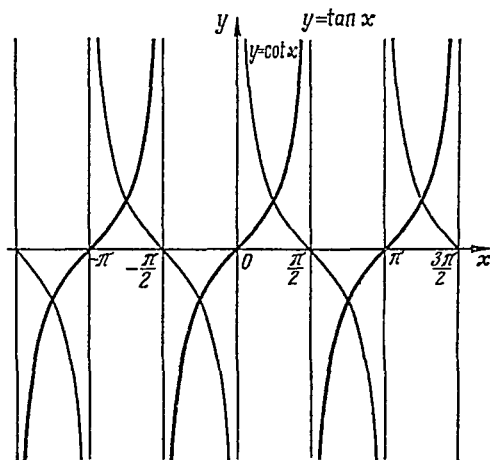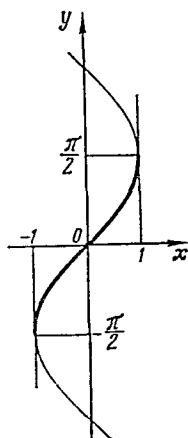
Fig. 47   Fig. 48

therefore (see Secs. 20-21) one usually considers its **principal branch** (the **principal value** of the arc sine) which is shown in Fig. 48 in heavy line; this branch is denoted as

$$y = \text{arc} \sin x, \quad -\frac{\pi}{2} \leqslant \text{arc} \sin x \leqslant \frac{\pi}{2}$$

and is a single-valued function. Other branches of the function have no special names.

The functions $y = \text{Arc} \cos x$ and $y = \text{Arc} \tan x$ can be investigated in a similar way and we leave this to the reader.

We should note in conclusion that we shall always deal with dimensionless (abstract) values of arc sin x. For example, we have

$$2^{\text{arc} \sin 1} = 2^{\frac{\pi}{2}} = 2^{1.57} = 2.97 *$$

Similarly, the values of the function $y = \sin x$ are taken for dimensionless values of x. We mean here that the sine of a number x

_____

* The last two equalities are approximate. If one intends to stress this fact one writes $2^{\frac{\pi}{2}} \approx 2^{1.57}$. We are not going to mention stipulations of this kind in the future.

is the sine of the angle of x radians. For instance, $\sin 1 = \sin 57°18' = = 0.8415$.

30. Empirical Formulas. We have already mentioned (see Sec. 13) that an experiment often results in a function $y = f(x)$ which we are interested in and represents the function in a tabular form (2). In such a case the problem of selecting an appropriate empirical formula for the function may arise. We usually begin with representing the values of the function on the graph paper or some other appropriate paper. Then we select a certain form of the formula we are going to use. If the form is not implied by general considerations we usually choose one of the functions described in Secs. 22-29 or a simple combination of such functions (a sum of power functions or of exponential functions and the like). In order to select such a formula in the best way one must know the graphs of these functions well. When selecting a function we must try to achieve the resemblance between the characteristic peculiarities of a sought-for function $\varphi(x)$ and of the function $f(x)$ under consideration. For example, if the physical meaning of the function indicates that $f(x)$ is even and $f(0) = 0$ then the function $\varphi(x)$ should also have these properties and so on.

Fig. 49

It sometimes turns out that we cannot find a single formula for the whole interval of x. Then it is necessary to divide the interval into several parts and select, for each of the parts, its own appropriate formula.

After the form of the formula has been chosen it is necessary to determine the values of parameters entering into the formula.

For example, suppose that after plotting the points we obtain the drawing shown in Fig. 49. If we have certain reasons to suspect that the experiment or the calculations of the values of the function could contain essential errors we must simply discard the points which fall out of the general form of the relationship described by the data represented in our drawing. For instance, the point P in Fig. 49 is a point of this kind. By the way, such points may sometimes indicate that certain important factors were not taken into account and then, of course, we must pay much attention to them.

The remaining points in Fig. 49 resemble a linear relation of the form $y = ax + b$. In order to determine the parameters a and b let us draw a straight line such that the experimental points should lie as close as possible to the line. This can easily be done by means of a transparent ruler. We apply the ruler to the drawing and then
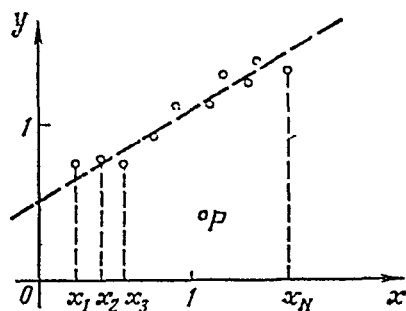
approximately find the sought-for position of the ruler. For example, the straight line drawn in Fig. 49 yields $b = 0.50$ and $a = \frac{\Delta y}{\Delta x} = 0.58$, i.e. $y = 0.58x + 0.50$.

The selection of a linear relation described above is comparatively simple. Therefore when choosing some other kind of functional relation one often tries to introduce new variables so that there should be a linear relation between the new variables and then to determine the parameters entering into the linear relation. We can apply this method only if there are no more than two such parameters since a linear function contains two parameters.

For example, let an experiment yield the following table of values:

x	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
y	0.00	0.01	0.03	0.08	0.17	0.29	0.45	0.66	0.91	1.22	1.57

We leave to the reader to represent the experimental points on the graph paper. The disposition of the points thus constructed resembles the graph of a power function of the form $y = ax^\alpha$. In
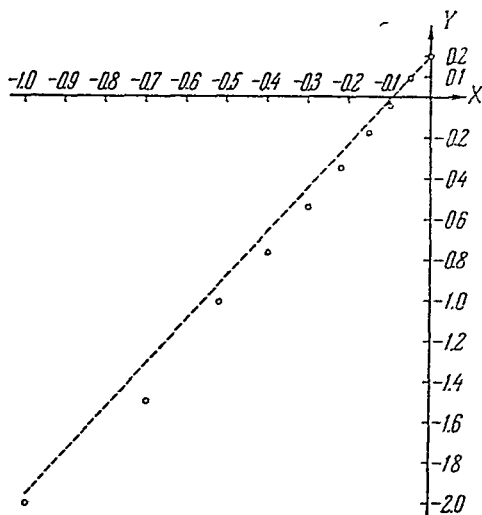
Fig. 50

order to determine the parameters a and α we take logarithms of both sides of the equality and denote $\log y = Y$, $\log x = X$ and $\log a = A$. Then we arrive at the equality $Y = \alpha X + A$ and thus

we see that there is a linear relation between the new variables. By means of a table of logarithms we compile the table of the values of the new variables:

X	—1.0	—0.70	—0.52	—0.40	—0.30	—0.22	—0.15	—0.10	—0.05	0.00
Y	—2	—1.5	—1.1	—0.77	—0.54	—0.35	—0.18	—0.041	0.086	0.196

The points thus obtained are lying close enough to the straight line drawn in Fig. 50. In drawing the straight line we should pay more attention to the last three points whose positions are determined with greater accuracy. Our construction yields the values $A = 0.196$ and $\alpha = 2.44$, that is $a = 1.57$. Hence we finally obtain $y = 1.57x^{2.44}$.

Some further rules and examples see in [51].

Plane Analytic Geometry

Analytic geometry is a branch of mathematics in which geometrical problems are investigated on the basis of the coordinate method by means of algebraic techniques.

§ 1. Plane Coordinates

1. Cartesian Coordinates. Cartesian coordinates are known from elementary mathematical courses and we have already used them (see Chapter I). Cartesian coordinates are called after R. Descartes. R. Descartes and P. Fermat (1601-1665) are the founders of the coordinate method.

Several points are depicted in Cartesian coordinates in Fig. 51. It should be noted that we take mutually perpendicular coordinate axes here and that the unit of length is the same for both axes. The origin of the coordinate system is placed at the point of intersection of the axes from which the distances along the axes are reckoned. (As it was mentioned in Sec. I.14, when we construct graphs, we can sometimes take different scales for the axes and change the position of the point the coordinates are reckoned from.) *Each point in the coordinate plane has certain uniquely determined coordinates and, conversely, to each ordered pair of coordinates x and y there corresponds a certain uniquely determined point of the plane.* This basic property makes it possible to consider the coordinates of points instead of the points themselves.
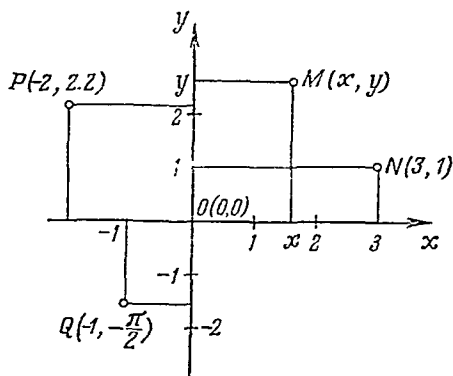
Fig. 51

The coordinate axes break the plane into the **quarters (quadrants)** which are numbered in the way shown in Fig. 52. Each of these quadrants is characterized by its specific combination of the signs of abscissas and ordinates; this is also shown in Fig. 52.

2. Some Simple Problems Concerning Cartesian Coordinates.

(1) *The distance between two given points.* Let the points M_1 (x_1, y_1) and M_2 (x_2, y_2) be given (i.e. their coordinates are known). It is required to determine the distance $d = M_1M_2$ (see Fig. 53). The
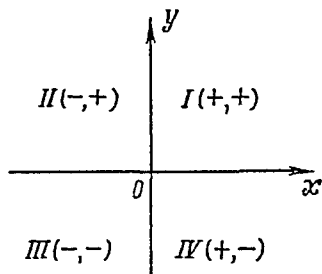
Fig. 52
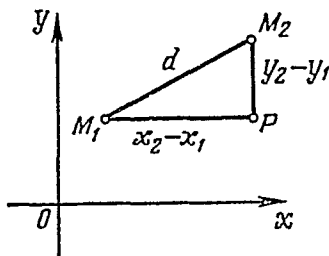
Fig. 53

formula for the distance is implied by Pythagoras' theorem applied to the rectangular triangle M_1M_2P. Thus we have $M_1M_2^2 = M_1P^2 + PM_2^2$, i.e. $d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$, or

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

This formula and all the following formulas hold for any two points M_1 and M_2 placed in an arbitrary manner in the coordinate plane.

(2) *Division of a line segment in a given ratio.* Let some points M_1 (x_1, y_1) and M_2 (x_2, y_2) be given. It is required to find a point M (x, y) lying on the segment M_1M_2 such that the ratio of division $\frac{M_1M}{MM_2}$ should be equal to λ where λ is a given number (see Fig. 54). The solution of the problem follows from the similarity of the triangles M_1PM and MQM_2 which implies $\frac{M_1P}{MQ} = \frac{M_1M}{MM_2} = \lambda$, i.e. $\frac{x - x_1}{x_2 - x} = \lambda$ and $x - x_1 = \lambda x_2 - \lambda x$. From the last relations we deduce

$$x = \frac{x_1 + \lambda x_2}{1 + \lambda}, \qquad y = \frac{y_1 + \lambda y_2}{1 + \lambda} \tag{2}$$

(the expression for y is obtained in like manner). In particular, in the case $\lambda = 1$, that is when the segment is halved, we have

$$x = \frac{x_1 + x_2}{2}, \qquad y = \frac{y_1 + y_2}{2}$$

(3) *Transformation of coordinates without changing the scale.* Suppose we have an "old" coordinate system x, y. Now let a "new" coordinate system x', y' be introduced. It is required to establish the relation-
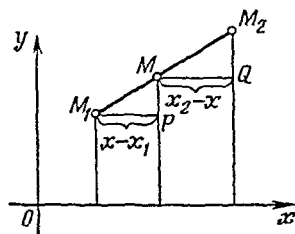
Fig. 54
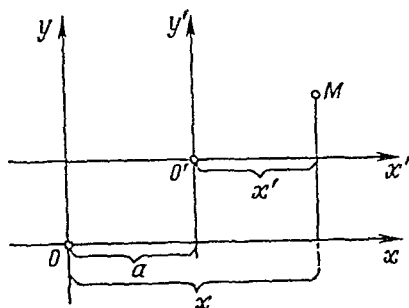
Fig. 55

ship between the old coordinates and the new ones. We shall consider the following three cases.

I. Let the new coordinates be the result of a translation of the original ones. Suppose the new origin of the coordinate system has
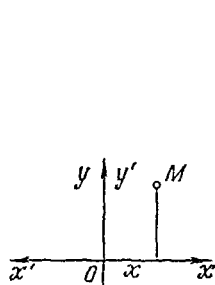
Fig. 56
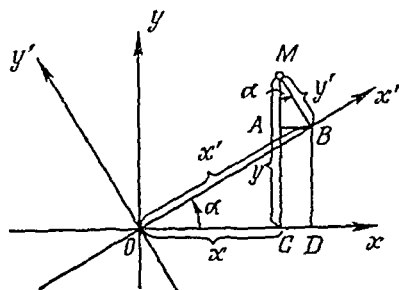
Fig. 57

the coordinates (a, b) with respect to the original coordinate system. Then, as it is implied by Fig. 55, we have

$$x = x' + a, \quad y = y' + b$$

II. Suppose the new coordinate axes are the mirror image of the old ones (for example, the mirror image about the y-axis). Then, as it is seen in Fig. 56, the relationship is

$$x = -x', \quad y = y' \tag{3}$$

III. Let now the new axes be the result of turning the old axes about the origin of the coordinate system through an angle α (see

Fig. 57). Then the equalities $OC = OD - AB$ and $CM = DB + AM$ imply

$$\left.\begin{array}{l} x = x' \cos \alpha - y' \sin \alpha \\ y = x' \sin \alpha + y' \cos \alpha \end{array}\right\} \tag{4}$$

The general case of passing from one Cartesian system to another is a mere combination of the above basic cases.

3. **Polar Coordinates.** Besides Cartesian coordinate systems we can construct many other different coordinate systems, that is there
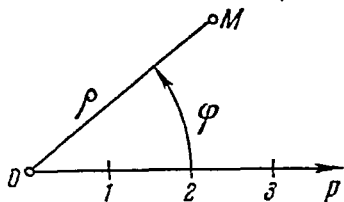
Fig. 58
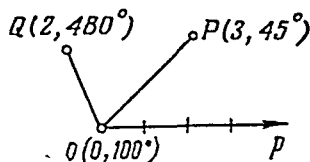
Fig. 59

are many ways to determine the position of a point in a plane by means of two numerical parameters (coordinates). An appropriate system should be chosen for each occasion, Cartesian systems being applied more often than others. In this section we shall consider only one of the systems, namely, the so-called polar coordinate system which is particularly convenient for investigating rotary motion. In order to construct polar coordinates we arbitrarily choose a pole O and a polar axis Op (see Fig. 58). Then the position of a point M can be characterized by its polar radius (radius-vector) ρ (i. e. the distance from O to M) and its polar angle (vectorial angle) φ (which

Fig. 60

is also called the phase or the amplitude of the point M). ρ and φ are called the polar coordinates of the point M. The vectorial angle φ is considered positive if it is reckoned in the positive direction (which is usually taken counterclockwise) and negative if otherwise. Several points with given polar coordinates are constructed in Fig. 59. In particular, we see that the pole has the zero radius-vector while its vectorial angle is undetermined and can be chosen in an arbitrary way. In order to describe all the positions a point can occupy in a plane it is sufficient to take the angle φ within the limits $-180° < \varphi \leqslant 180°$ but it may sometimes be convenient to consider the values of φ which fall outside the interval. The
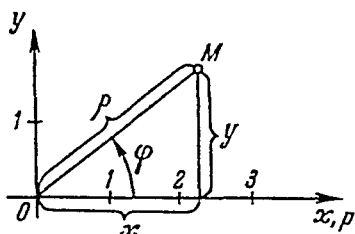
addition of 360° to the vectorial angle of a point does not change its position.

The relationship between the Cartesian coordinates and the polar coordinates of a point in case the systems are placed as it is shown in Fig. 60 has the form

$$x = \rho \cos \varphi, \quad y = \rho \sin \varphi \quad \text{and, conversely,}$$

$$\rho = \sqrt{x^2 + y^2}, \quad \tan \varphi = \frac{y}{x} \tag{5}$$

§ 2. Curves in Plane

4. Equation of a Curve in Cartesian Coordinates. As it was shown in Sec. I.20, an equation of the form

$$F(x, y) = 0 \tag{6}$$

defines a curve (L) in the x, y-plane (i.e. in the plane where the coordinate system is taken). The curve is the locus of all the points whose coordinates satisfy equation (6). The relation of form (6) is called the equation of the curve (L).

Fig. 61

Conversely, if a curve (L) is given in the x, y-plane then its geometrical properties can be expressed in terms of some analytic relations between the coordinates of its points, and thus an equation of the curve (L) is obtained in form (6). (It should be taken into account of course that every equation can be rewritten in different equivalent forms.) Consequently, the coordinate method enables us to consider the equations of curves instead of the curves themselves. Thus, geometrical problems can be reduced to algebraic ones, and the latter, as a rule, can be solved in a simpler and more uniform way than the former. For example, in order to verify whether a curve with an equation of form (6) (or, as we simply say, "the curve $F(x, y) = 0$") passes through a point (a, b) it is sufficient to substitute the coordinates of the point into the equation of the curve and verify whether the equation is satisfied, i.e. whether we have $F(a, b) = 0$.
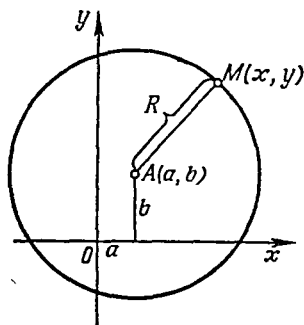
As an example, let us deduce the equation of a circle (see Fig. 61). Let the centre A of the circle have the coordinates (a, b) and let $M(x, y)$ be an arbitrary (moving, current) point of the circle. Then the basic property characterizing a circle can be written as $AM = R$ where R is the radius of the circle. Now, applying formula (1) for

the distance between two points we obtain

$$\sqrt{(x-a)^2 + (y-b)^2} = R$$

or, squaring, we derive

$$(x-a)^2 + (y-b)^2 = R^2$$

This relation is the one which is satisfied by the coordinates of all the points of the given circle and at the same time only the points belonging to the circle may satisfy this relation. Thus, this relation is an equation of the circle and a, b and R entering into the equation are some fixed numbers (i.e. parameters which characterize the position and the size of the circle) whereas x and y are the current coordinates of a variable point of the circle.

Now, contrary to the previous example, let an equation be originally given, for example, the equation

$$x^2 + y^2 - 3x + 4y - 1 = 0 \qquad (7)$$

Transforming the equation and completing the squares we obtain

$$\left(x - \frac{3}{2}\right)^2 - \left(\frac{3}{2}\right)^2 + (y+2)^2 - 2^2 - 1 = 0,$$

$$\left(x - \frac{3}{2}\right)^2 + (y+2)^2 - \frac{29}{4} = 0$$

Consequently the given equation is the equation of a circle with centre at the point $(1.5, -2)$ and of radius $\sqrt{\dfrac{29}{4}} = 2.69$.

If two curves with the equations $F_1(x, y) = 0$ and $F_2(x, y) = 0$ are given we can pose the problem of finding the point of intersection of the curves. The sought-for point of intersection must belong to both lines simultaneously and therefore its coordinates x and y must satisfy the equations of both lines. Hence, in order to determine the coordinates we must solve the following system of equations:

$$\left.\begin{array}{l} F_1(x, y) = 0 \\ F_2(x, y) = 0 \end{array}\right\} \qquad (8)$$

Such a system of equations can have a number of distinct solutions and this number corresponds to the number of possible points of intersection. Of course, a solution is understood as a pair of certain values x and y satisfying (8).

For example, let it be necessary to determine the point of intersection of circle (7) with the straight line $y = x + b$ where b is a constant. To do this we should solve the system of equations

$$\left.\begin{array}{l} x^2 + y^2 - 3x + 4y - 1 = 0 \\ y = x + b \end{array}\right\}$$

Expressing y in terms of x from the second equation, substituting the expression thus obtained for y into the first equation, removing brackets and solving the quadratic equation in x we obtain after some transformations

$$x_1 = \frac{-1-2b+\sqrt{9-28b-4b^2}}{4}, \quad y_1 = \frac{-1+2b+\sqrt{9-28b-4b^2}}{4};$$

$$x_2 = \frac{-1-2b-\sqrt{9-28b-4b^2}}{4}, \quad y_2 = \frac{-1+2b-\sqrt{9-28b-4b^2}}{4}$$
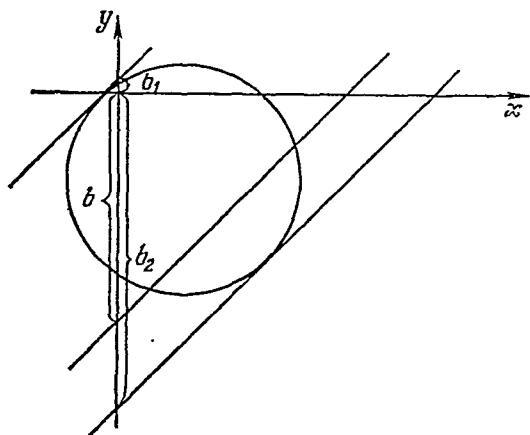
Fig. 62

Let us find for what values of b both points of intersection coincide. This will happen in case the expression under the radical sign vanishes which implies $b_{1,2} = \frac{-7 \pm \sqrt{58}}{2}$, i.e. $b_1 = 0.31$ and $b_2 = -7.31$. The straight line $y = x + b$, as it is shown in Fig. 62, is tangent to the circle for these values of b. If $b_1 < b < b_2$ then there are two distinct points of intersection: (x_1, y_1) and (x_2, y_2). The straight line does not intersect the circle at all when b lies outside the above interval (in this case the expression under the radical sign is negative).

It may be noted that in many different problems the coincidence of the points of intersection whose coordinates are found from a system of type (8) usually indicates that the two given curves are tangent to each other at this common point, that is they have a common tangent line at the point.

5. Equation of a Curve in Polar Coordinates. An equation connecting the plane coordinates of points taken with respect to any

given coordinate system defines some curve in the coordinate plane (with some exceptions to this rule which will be discussed in Sec. 8). In particular, let us consider a polar coordinate system. We suppose that the equation is solved for ρ, i.e. it has the form

$$\rho = f(\varphi) \qquad (9)$$

Making φ assume all the possible numerical values and finding the corresponding values of ρ we receive the locus of these points which
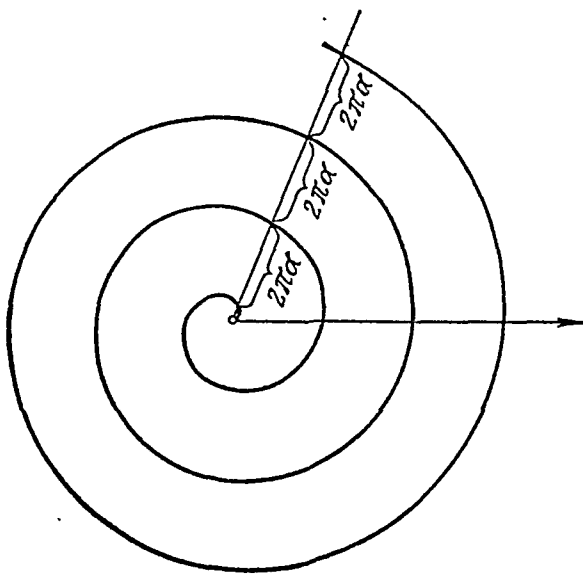
Fig. 63

form a curve in the plane, that is the graph of function (9) in the polar coordinates.

Let us take two examples. The graph of a linear relation of the form $\rho = a\varphi + b$ is shown in Fig. 63. This curve is called the **spiral of Archimedes.** It can be obtained by the superposition of a uniform rotary motion of the radius-vector and a uniform rectilinear motion along the radius. Indeed, if

$$\rho = vt + b \quad \text{and} \quad \varphi = \omega t \quad \text{then} \quad \rho = \frac{v}{\omega}\varphi + b$$

Thus we see that the graph of one and the same function (a linear function in this particular case) regarded with respect to a Cartesian coordinate system and to a polar coordinate system has quite different forms (see Sec. I.22).

The graph of the exponential function $\rho = e^{h\varphi}$ in polar coordinates is depicted in Fig. 64. This curve is the so-called **logarithmic**

spiral. The curve infinitely winds round the pole (and tends to the pole as $\varphi \to -\infty$) but never reaches it.

The logarithmic spiral has some interesting properties. For example, if we perform a **similarity transformation**, that is if we expand the spiral uniformly in all directions, with the **similarity coefficient** m (i.e. all the linear sizes are increased m times) we shall obtain a new curve with the equation $\rho = me^{h\varphi}$. But $\rho = me^{h\varphi} = e^{k\left(\varphi + \frac{\ln m}{k}\right)} = e^{k(\varphi + \alpha)}$ where $\alpha = \dfrac{\ln m}{k}$ and therefore the result would be as
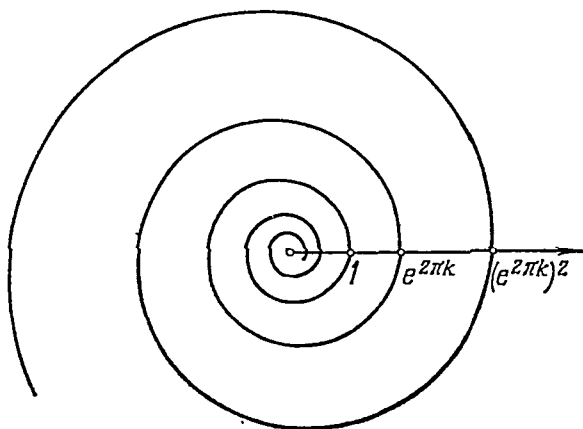
Fig. 64

if we turned the original spiral about the pole through the angle of α radians in the clockwise direction. Indeed, generally the graph of a function $\rho = f(\varphi + \alpha)$ is obtained from the graph of the function $\rho = f(\varphi)$ by rotating the latter about the pole through the angle α in the negative (i.e. clockwise) direction. (Why is it so?) Consequently, the logarithmic spiral is similar to itself with any arbitrary similarity coefficient. Only straight lines possess this property among all the other curves in a plane.

In conclusion, let us discuss the notion of **coordinate curves**. A curve is called a coordinate curve if one of the coordinates remains constant along the curve. The coordinate curves of a Cartesian coordinate system are two families of straight lines parallel to the coordinate axes. The curves $\rho = \text{const}$ which are concentric circles with centre at the pole form a family of coordinate curves in a polar coordinate system. The second family of coordinate curves in a polar system are the curves $\varphi = \text{const}$, that is the family of half-lines issued from the pole (see Fig. 65).

6. Parametric Representation of Curves and Functions. There are some cases when both coordinates of a point (for instance, the Cartesian coordinates) belonging to the coordinate plane are represented as certain functions of a third variable which we denote by t:

$$x = \varphi(t), \quad y = \psi(t) \tag{10}$$

This third variable serves as a parameter determining the position of a point having the coordinates (x, y) in the coordinate plane. When t varies the corresponding point moves in the plane describing some curve (L) (see Fig. 66). In this case we say that *the curve (L) is represented in parametric form* (10). We have such a situation when,
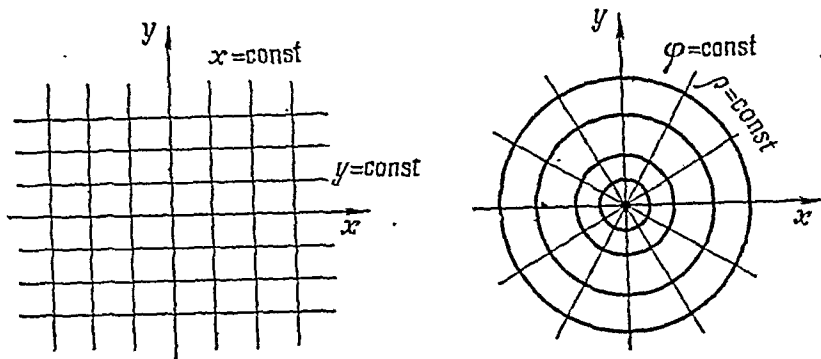
Fig. 65

for example, we investigate a motion of a point in the plane. In this case t is time, formulas (10) determine the law of motion and the curve (L) is called the **trajectory** of motion. In other problems of this kind a parameter entering into equations may have some other physical or geometrical meaning but even then it is usually convenient to regard it as if it were time. It should be noted that one and the same curve (L) can be represented by many different forms of equations (10) since there can be different laws of motion along one and the same trajectory (for instance, students walking from a bus stop to their college can come 22 minutes or 2 minutes before the lecture begins).

In order to pass from equations of a curve given in form (10) to an equation of general form (6) we must eliminate the parameter from both equations (10). For example, we can express t in terms of x from the first equation and then substitute the result into the second equation. Of course, we can use any other procedure which eliminates t. But this is not always possible and besides we can sometimes find it inexpedient to do so. Therefore we often retain the parametric form of representation.

Equations (10) define a certain functional relation $y\,(x)$. Indeed, if we, for example, make x take on some value then the first equation defines some value of t (of course, there may be more than one such value) and the second equation defines some value (or several values) of y. We see that the function $y\,(x)$ turns out to be represented in a parametric form and the curve (L) is its graph.

Example 1. Let us consider the motion of a shell fired with the initial velocity v_0 at an angle α to the horizon (see Fig. 67). We shall disregard air resistance, sphericity of the earth and the earth's
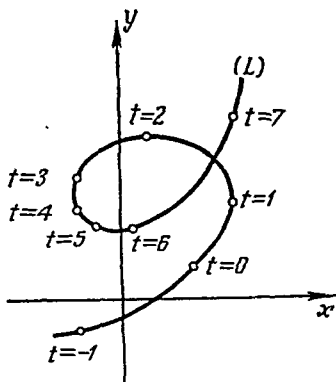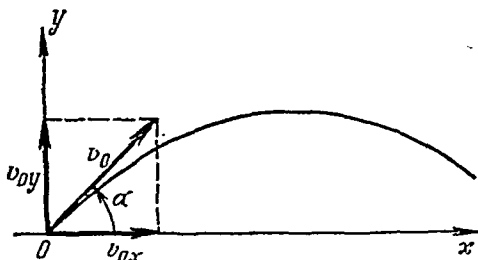
Fig. 66   Fig. 67

rotation. Then the horizontal component of the velocity must be permanently constant and equal to $v_{0x} = v_0 \cos \alpha$ whereas the vertical component of the velocity changes all the time. The vertical motion has the constant acceleration of gravity g and therefore the distance passed along the vertical differs from the vertical displacement which corresponds to the constant speed $v_{0y} = v_0 \sin \alpha$ by $\frac{gt^2}{2}$ (this fact is well known from mechanics). Thus, we obtain the law of motion in the x, y-plane: $x = (v_0 \cos \alpha)\,t$ and $y = (v_0 \sin \alpha)\,t - \frac{gt^2}{2}$. This is the law of motion that was sought for and at the same time it defines the trajectory of motion in the parametric form. Eliminating t we deduce

$$y = (\tan \alpha)\,x - \frac{gx^2}{2v_0^2 \cos^2 \alpha} \qquad (11)$$

The dependence $y\,(x)$ being a quadratic one, the trajectory is a parabola (see Sec. I.23).

Example 2. Let us now consider the trajectory of a point of a circle rolling upon a straight line without sliding (the dotted line in

Fig. 68 shows the position of the circle at the initial moment $t = 0$ whereas the continuous line indicates some moving position). We shall regard the angle of rotation of the circle (denoted by ψ) as a parameter. Then

$$
\left.
\begin{aligned}
x = OQ &= OT - QT = \smile MT - MP = R\psi - \\
&- R \sin \psi = R (\psi - \sin \psi) \\
y = QM &= TN - PN = R - R \cos \psi = \\
&= R (1 - \cos \psi)
\end{aligned}
\right\}
\tag{12}
$$

where the line segment OT is equal to the arc length of $\smile MT$ according to the condition of the absence of sliding. Hence, we have obtain-
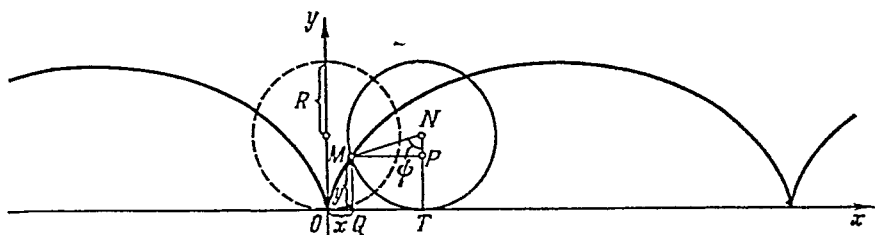
Fig. 68

ed the parametric equations of a curve which is called the cycloid. The curve is infinite and has cusps which are seen in Fig. 68.

The cycloid is one of the simplest curves belonging to the class of roulettes; a **roulette** is the path in a fixed plane of any point in
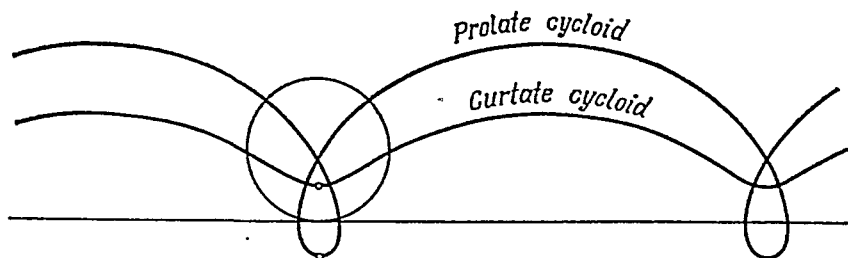
Fig. 69

a moving coincident plane when a given curve in the latter plane rolls without sliding on a given curve in the former. The so-called **curtate** and **prolate** cycloids (see Fig. 69) which are described by a point rigidly connected with the plane of a circle and lying. respectively, inside or outside the circle when the latter is rolling

upon a straight line give further examples of roulettes. By the way, the prolate cycloid, as it is seen in Fig. 69, has nodal points. Other examples are the **hypocycloid** and the **epicycloid** which are described by a point of a circle rolling upon the other circle inside or outside it, respectively (see Fig. 70). A hypocycloid with the ratio of the radii of the circles equal to $1 : 4$ is called the **astroid**. In case this
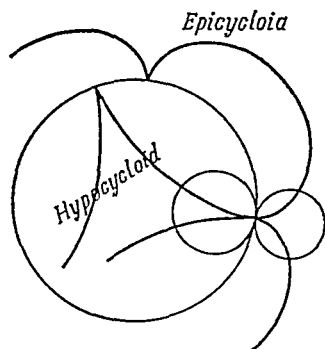
Fig. 70
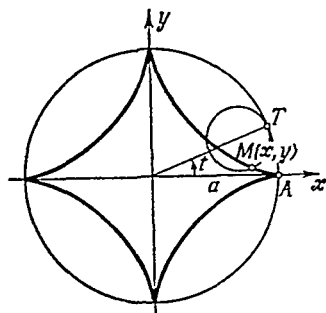
Fig. 71
Astroid. Show that the equality $\cup\ TM = \cup\ TA$ implies the equations $x = a \cos^3 t$, $y = a \sin^3 t$

ratio is $1 : 2$ the hypocycloid turns into a straight line segment. The **cardioid** is an epicycloid with the ratio of the radii $1 : 1$. These curves are depicted, respectively, in Figs. 71 and 72, where the equations of the curves are also put down. The deduction of these equations is left to the reader. All these curves are important for applications in the theory of mechanisms.

7. **Algebraic Curves.** If the equation of a curve (L) has the form

$$P(x, y) = 0 \qquad (13)$$

in Cartesian coordinates x and y where P is a polynomial of degree n the curve (L) is said to be an algebraic curve of the nth order. Non-algebraic curves are called transcendental. For example (see Sec. I.22), the graph of a linear function, that is a straight line, is an algebraic curve of the first order, a quadratic parabola and a circle are algebraic curves of the second order, the cubic parabola and the semicubical parabola [the latter curve has the equation $y = x^{\frac{2}{3}}$ which should be rewritten in the form $y^3 - x^2 = 0$ in order to obtain an equation of form (13)] are algebraic curves of the third order. On the other hand, the sinusoid, the tangent curve and the graph of an exponential function are transcendental curves.

As an example of a curve of the fourth order let us consider Cassinian ovals introduced by the Italian astronomer G. D. Cassini (1625-1712). Let us take two points F_1 and F_2 in a plane and form the product of their distances from a point M in the plane. The locus of all points M in the plane for which the product is a constant value is a Cassinian oval. Let us introduce Cartesian coordinate
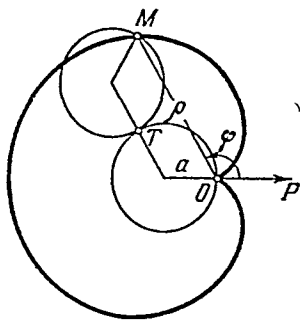
Fig. 72

Cardioid. Show that the equa-
lity $\cup TM = \cup TO$ implies
the equation $\rho = 2a(1 - \cos \varphi)$
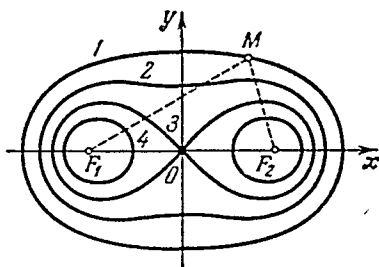
Fig. 73

axes in the way shown in Fig. 73. Denoting the coordinates of the points F_1, F_2 and M as $F_1(-a, 0)$, $F_2(a, 0)$ and $M(x, y)$, respectively, we obtain, by formula (1), the equation

$$(\sqrt{(x + a)^2 + y^2}) \cdot (\sqrt{(x - a)^2 + y^2}) = b^2$$

where b^2 is the given constant value of the product. The last equation implies the final equation

$$(x^2 + y^2)^2 - 2a^2(x^2 - y^2) = b^4 - a^4$$

which can be easily deduced. Here we have an important special case when $b = a$ and the curve is ∞-shaped and called the lemniscate. The lemniscate was discovered in 1694 by the Swiss mathematician Jakob Bernoulli (1654-1705). Passing to the polar coordinates by means of formulas (5) we deduce the polar equation of the lemniscate $\rho^2 = 2a^2 \cos 2\varphi$. In case $0 < b < a$ a Cassinian oval consists of two separate parts.

It should be noted that when we speak about the degree of an algebraic curve we always mean an equation in Cartesian coordinates. For example, the equation of the spiral of Archimedes (see Sec. 5) has the first degree in the polar coordinates but if we rewrite it in the Cartesian coordinates we get $\sqrt{x^2 + y^2} = a \arctan \frac{y}{x} + b$

and therefore we see that the spiral is a transcendental curve. Spirals of all other types are also transcendental and, generally, it turns out that all infinite curves possessing a periodicity property are also transcendental.

The degree of an algebraic curve does not change if we replace a given Cartesian coordinate system by another one.

Actually, for example, if we perform a parallel translation of a coordinate system (see Sec. 2) then equation of a curve (13) turns into

$$P(x' + a, \ y' + b) = 0$$

The degree of a polynomial which is obtained after removing brackets and collecting similar terms cannot become higher than the original one (obviously, terms of higher degree cannot appear here in this case). One may think that after collecting terms the degree of the polynomial could decrease in case the terms of higher degree mutually cancel out. But this cannot occur since otherwise the degree of the polynomial should increase under the inverse transition from x', y' to x, y and this, as it was shown above, is impossible. An analogous situation takes place for all other types of transformation of Cartesian coordinates.

A change of a coordinate system for an immovable curve is equivalent to the motion of the curve (in the opposite direction) as the axes of coordinates remain immovable and therefore *the degree of an algebraic curve is invariant (unalterable) when the curve moves as a whole.*

A quantity or, in general, an object which does not change under some transformations is called an **invariant** of these transformations (or an invariant with respect to the transformations). For example, the area is an invariant of motions and angles are invariants not only of motions but of similarity transformations as well. Thus, the order of an algebraic curve is also an invariant of motions. This very important concept of an invariant was unfamiliar to an owner of a garden who was painting the fence. Knowing that he did not have enough paint he was doing the job extremely fast so as to finish it before the paint was used up.

8. Singular Cases. There are some equations of the form $F(x, \ y) = 0$ which define in the x, y-plane such a set of points that it would hardly be possible to call it a line or a curve. We shall illustrate such cases by some examples.

The equation $x^2 + y^2 + 1 = 0$ has no points in the coordinate plane which satisfy it since the left-hand side is positive for all x and y. An equation of this kind is said to describe an **imaginary curve** since if we use complex numbers we can put, for instance, $x = i$, $y = 0$ etc. But this term does not change the fact that such a curve does not exist as a real curve in the plane.

Only one point $x = 0$, $y = 0$ (the origin of the coordinate system) satisfies the equation

$$x^2 + y^2 = 0 \qquad \square \qquad (14)$$

Comparing this equation with the equation

$$x^2 + y^2 - R^2 = 0 \qquad (15)$$

(see Sec. 4) we may consider equation (14) as if it defined a circle with zero radius. Generally, if an object depends on parameters and if it changes considerably and loses some of its essential properties for certain values of the parameters we usually say that the object degenerates for these values. In the above case circle (15) depends on the parameter R. It degenerates into point (14) for $R = 0$ and loses its main property, i.e. the property to be a curve.

The equation of an algebraic curve of the second order of the form

$$y^2 - x^2 = 0 \qquad (16)$$

can be rewritten as

$$(y - x)(y + x) = 0$$

Fig. 74

But a product is equal to zero if and only if at least one of its factors equals zero. Therefore, either $y - x = 0$, i. e. $y = x$, or $y + x = 0$, i. e. $y = -x$ (see Fig. 74). Each of these equations determines a straight line in the x, y-plane. Consequently, a point satisfying equation (16) lies either on the first straight line or on the second one. Thus, a curve defined by equation (16) is nothing but a pair of straight lines or, as we say, it disintegrates into a pair of straight lines. Hence, we see that the algebraic curve of the second order disintegrated into two straight lines.

But we can by no means regard a hyperbola as a curve which disintegrates (Sec. I.25). In this case we have a curve consisting of two components (branches) and each of these components is a half of the hyperbola but both branches are described by one and the same equation. As for the case of equation (16), each of the straight lines obtained above has its own self-dependent equation. It appears quite clear that in this way we can artificially unite two arbitrary curves. If these curves have equations $F_1(x, y) = 0$ and $F_2(x, y) = 0$ then it is sufficient to take the equation

$$F_1(x, y) \cdot F_2(x, y) = 0$$

We cannot regard as a disintegration of an algebraic curve the decomposition of the parabola (shown in Fig. 23) into its upper and
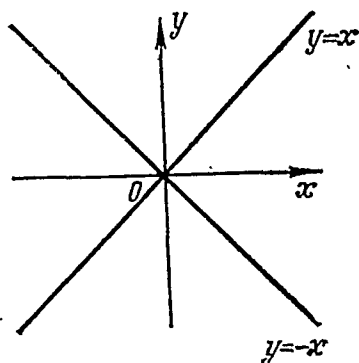
lower halves according to the formula

$$y^2 - x = (y - \sqrt{x})(y + \sqrt{x}) = 0 \qquad (17)$$

from which we obtain $y_{1,2} = \pm\sqrt{x}$ instead of $y^2 - x = 0$. Here the key point is that we have a disintegration when a polynomial on the left-hand side of equation (13) is factored into polynomials whereas the factors entering into the right-hand side of (17) are not polynomials (see Sec. I.17).

§ 3. First-Order and Second-Order Algebraic Curves

9. Curves of the First Order. As is shown in Sec. 7, in order to obtain an algebraic curve of the first order we must take a polynomial of the first degree and equate it to zero. Such a polynomial may contain only terms of the first degree and an absolute (constant) term. Therefore the general form of an equation of a curve of the first order is the following:

$$Ax + By + C = 0 \qquad (18)$$

There may occur two cases. If $B \neq 0$ then dividing the equation by B and denoting

$$-\frac{A}{B} = k, \qquad -\frac{C}{B} = b \qquad (19)$$

we receive

$$y = kx + b \qquad (20)$$

We have shown (see Sec. I.22) that this is the equation of a straight line (we had a instead of k but this fact is of no importance). This line is depicted in Fig. 75. In case $B = 0$ we divide the equation by A and denote $-\frac{C}{A} = a$. This yields an equation of the form $x = a$ which is the equation of a straight line parallel to the y-axis. It should be noted that for such a line the slope $k = \tan\frac{\pi}{2} = \pm\infty$ which also follows from relation (19) but in this case the equation cannot be written in form (20).

Thus, *the curves of the first order are straight lines.*

Let us consider several simple problems involving equations of straight lines.

1. Let it be required to construct a straight line with the given slope k passing through the given point (x_1, y_1). When we say "to draw" or "to construct" a straight line we mean, of course, in terms of analytic geometry, to write down its equation. The sought-for equation has form (20) but b entering into the equation is unknown.

But the straight line passing through the given point,[r] the coordinates of the point must satisfy the equation of the line: $y_1 = kx_1 + [b$. Subtracting and thus eliminating b we arrive at the sought-for equation:

$$y - y_1 = k(x - x_1) \tag{21}$$

Making k change and take all the possible values we obtain the pencil of all possible straight lines passing through the point (x_1, y_1). We can also put $k = \pm\infty$ and thus obtain the vertical straight
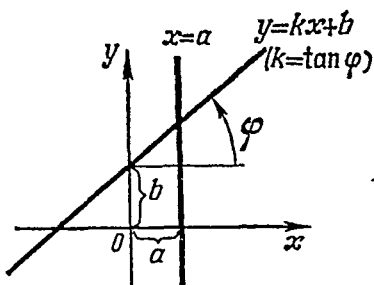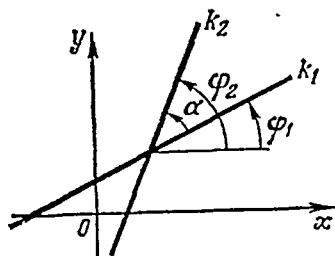
Fig. 75　　　　　Fig. 76

line. But to do this we should divide both sides by k beforehand; then after the substitution of $k = \pm\infty$ we simply obtain $0 = x - x_1$. i.e. $x = x_1$. Similar precautions should be taken in other problems involving infinite values of parameters.

2. Let it be required to draw a straight line through two given points (x_1, y_1) and (x_2, y_2). The equation of the desired straight line has form (21) but now k is unknown. But the condition that the straight line should pass through the second point implies $y_2 - y_1 = k(x_2 - x_1)$. Performing the division we thus eliminate k and receive

$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1} \tag{22}$$

In this equation and in equation (21) x and y are current coordinates of a moving (variable) point of the sought-for straight line.

3. Let it be required to determine the angle formed by two straight lines with given slopes k_1 and k_2. The solution is implied by Fig. 76:

$$\tan \alpha = \tan(\varphi_2 - \varphi_1) = \frac{\tan \varphi_2 - \tan \varphi_1}{1 + \tan \varphi_1 \tan \varphi_2} = \frac{k_2 - k_1}{1 + k_1 k_2} \tag{23}$$

4. The condition for the parallelism of two straight lines is obvious: $k_1 = k_2$.

5. The condition for the perpendicularity of two straight lines follows from problem 3: we have $\alpha = \dfrac{\pi}{2}$ for mutually perpendicular

straight lines but $\tan \frac{\pi}{2} = \pm\infty$ and therefore $1 + k_1 k_2 = 0$ or $k_2 = -\frac{1}{k_1}$.

10. **Ellipse.** *An ellipse is the locus of all points in a plane for which the sum of their distances from two fixed points in that plane* (these points are called the **focuses of the ellipse**) *is a constant quantity.* This definition enables us to use the method of drawing an ellipse with the help of a taut thread illustrated in Fig. 77. This procedure
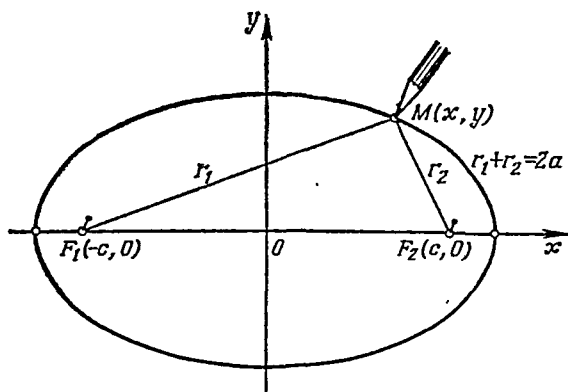
Fig. 77

allows us to visualize the form of an ellipse: the ellipse is a closed convex curve possessing two symmetry axes (called the **principal axes of the ellipse**) and having the centre of symmetry O (called the **centre of the ellipse**).

In order to deduce the equation of an ellipse in the simplest form let us place the coordinate axes in the way shown in Fig. 77 and denote $F_1 F_2 = 2c$ and $r_1 + r_2 = 2a$ where c and a are constants. Then on the basis of formula (1) we can write, according to the definition of an ellipse, $\sqrt{(x + c)^2 + y^2} + \sqrt{(x - c)^2 + y^2} = 2a$ from which we obtain, in succession,

$$\sqrt{(x+c)^2+y^2} = 2a - \sqrt{(x-c)^2+y^2}.$$

$$(x+c)^2+y^2 = 4a^2 - 4a\sqrt{(x-c)^2+y^2} + (x-c)^2 + y^2,$$

$$a\sqrt{(x-c)^2+y^2} = a^2 - cx, \quad a^2[(x-c)^2+y^2] = a^4 - 2a^2cx + c^2x^2,$$

$$x^2(a^2-c^2) + a^2y^2 = a^2(a^2-c^2) \tag{24}$$

It is seen, from the triangle $F_1 M F_2$, that $2a > 2c$, that is $a^2 - c^2 > 0$. Denote for brevity $a^2 - c^2 = b^2$. Then the last of rela-

tions (24) implies the so-called canonical form of the equation of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \qquad (25)$$

This equation again shows that the coordinate axes serve as the symmetry axes of the ellipse because if a point (p, q) satisfies equation (25) then the points $(-p, q)$, $(-p, -q)$ and $(p, -q)$ also satisfy it (see Fig. 78).
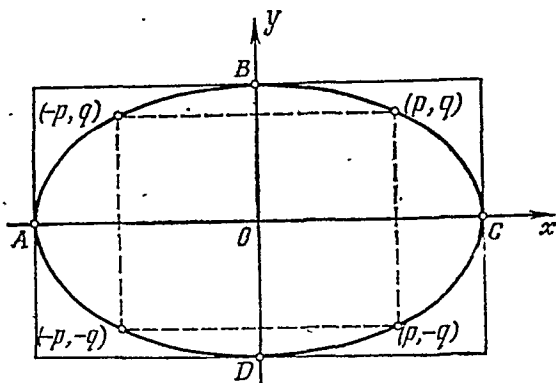
Fig. 78

Putting $y = 0$ we get $x = \pm a$ and, similarly, $x = 0$ yields $y = \pm b$. Hence, a and b are, respectively, the lengths of the semi-major axis and the semi-minor axis of the ellipse (see Fig. 78: $AO = OC = a$ and $DO = OB = b$). Besides, each summand entering into the left-hand side of (25) cannot be greater than unity and therefore $|x| \leqslant a$ and $|y| \leqslant b$. Consequently, the whole ellipse lies inside the rectangle depicted in Fig. 78. The points A, B, C and D at which the ellipse intersects its symmetry axes are called the vertices of the ellipse. An ellipse has four vertices.

The ratio $\varepsilon = \frac{c}{a} = \sqrt{\frac{a^2 - b^2}{a}}$, $0 < \varepsilon < 1$, is called the eccentricity of the ellipse. This is a dimensionless quantity which does not change under a similarity transformation of the ellipse when all its sizes are increased k times since $\frac{kc}{ka} = \frac{c}{a}$.

The eccentricity of an ellipse characterizes its form (its "elongation") but not its sizes. Several ellipses are depicted in Fig. 79; they all have the fixed length $2a$ of their major axes whereas their eccentricity ε varies, $c = \varepsilon a$ and $b = a\sqrt{1 - \varepsilon^2}$. This enables us

to see how the eccentricity affects the form of the ellipse: the focuses draw together and the minor axis tends to the major one in its length
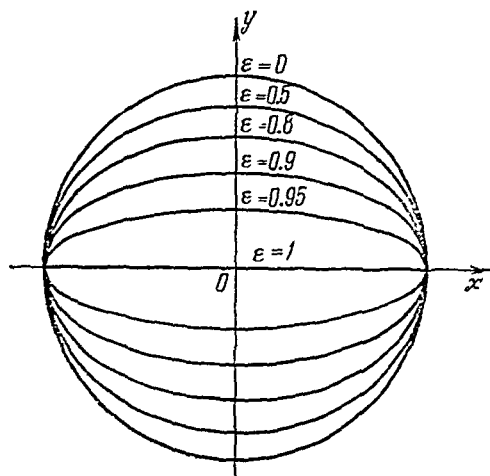
Fig. 79

as ε decreases. Passing to the limit as $\varepsilon \to 0$ we have $\varepsilon = 0$, $c = 0$ and $b = a$, that is we obtain a circle. Consequently. a circle may be regarded as a singular (limiting) case of an ellipse whose focuses merge and coincide with the centre of the circle; in this case the eccentricity is equal to zero. On the contrary, if ε approaches 1 the ellipse becomes more and more elongated and it degenerates into a straight line segment in the limiting process.
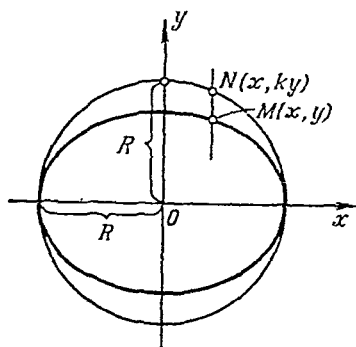
Fig. 80

An ellipse can be obtained by a uniform contraction of a circle in a certain direction. Indeed, let us, for instance, consider the uniform contraction towards the x-axis when all the sizes in the direction of the y-axis decrease k times (see Fig. 80). If a point $M(x, y)$ lies on the curve obtained as a result of the contraction then the point $N(x, ky)$ must belong to the circle. This implies $x^2 +$

$+ (ky)^2 = R^2$ or $\dfrac{x^2}{R^2} + \dfrac{y^2}{\left(\dfrac{R}{k}\right)^2} = 1$, i.e. we get an ellipse with the

semi-axes R and $\dfrac{R}{k}$.

We can easily deduce now the parametric equations of an ellipse using the above proved property. In fact, the equations $x = R \cos t$ and $y = R \sin t$ $(0 \leqslant t \leqslant 2\pi)$ define the *circle of radius R* (R is given) with centre at the origin of the coordinate system. (Verify
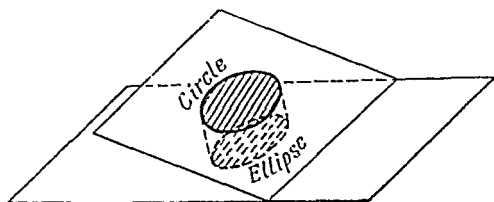
Fig. 81

this!) Performing the contraction we obtain $x = R \cos t$ and $y = \dfrac{R \sin t}{k}$. If now we introduce the semi-axes $a = R$ and $b = \dfrac{R}{k}$ we finally derive the parametric equations of the ellipse:

$$x = a \cos t, \quad y = b \sin t \quad (0 \leqslant t \leqslant 2\pi) \cdot \qquad (26)$$

We shall show in Sec. XI.6 that a uniform contraction of an ellipse again yields an ellipse.

As is known, an orthogonal projection of a plane figure results in a uniform contraction of the figure and therefore the orthogonal projection of a circle yields an ellipse (see Fig. 81).

We again obtain an ellipse when considering the section of a right circular cylinder or of a cone by a plane. Fig. 82 shows such a section for the case of a cylinder. In order to prove that the section represents an ellipse we inscribe two spheres into the cylinder in such a way that they should touch the plane at the points F_1 and F_2. As is known, two tangents drawn to a sphere from a common point are equal. Therefore, we can write, for any point M belonging to the section, $MF_1 + MF_2 = MN_1 + MN_2 = N_1N_2 = \text{const}$ (see Fig. 82). This relation implies that our assertion is true. The corresponding construction for the case of a cone is analogous. The properties of an ellipse are widely used in drawing.

11. Hyperbola. We have already dealt with a hyperbola (see Sec. I.25). But let us now forget about it for a while; later on in Sec. 13 we shall establish the connection between Secs. 11 and I.25. Here we shall give a new definition of a hyperbola: *a hyperbola is*

the locus of all points in a plane for which the difference between their distances to two given points (called the focuses of the hyperbola) *is a constant quantity.* Choosing the coordinate axes as is shown in Fig. 83 and denoting $F_1F_2 = 2c$ and $r_1 - r_2 = \pm 2a$ we obtain the equation of the hyperbola in the form $\sqrt{(x+c)^2 + y^2} - \sqrt{(x-c)^2 + y^2} = \pm 2a$. Now carrying out some transformations similar to those in Sec. 10 we arrive at the same relation (24). (Check it up!) But *in this case the triangle F_1MF_2 implies $2a < 2c$ and*
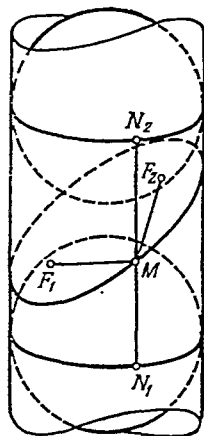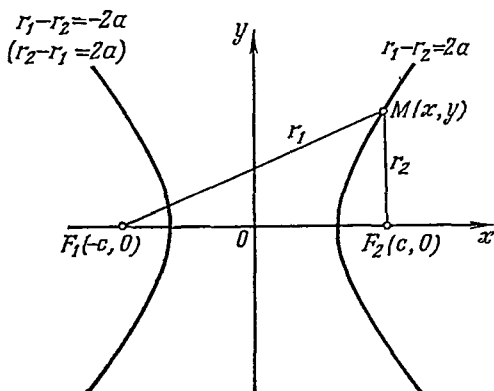
Fig. 82   Fig. 83

therefore we cannot denote $a^2 - c^2 = b^2$ as it was done in Sec. 10. (Why is it so?) *Therefore we denote $a^2 - c^2 = -b^2$ which is permissible.* Then we deduce from (24) the relation

$$-b^2x^2 + a^2y^2 = -a^2b^2$$

and finally get the canonical form of the equation of the hyperbola:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \tag{27}$$

This equation shows that a hyperbola has two symmetry axes (its **principal axes**) and a centre of symmetry (the **centre of the hyperbola**). Putting $y = 0$ we get $x = \pm a$ and putting $x = 0$ we obtain $y = \pm ib$. Consequently, the x-axis intersects the hyperbola at two points (which are the **vertices of the hyperbola**). This axis is called the **transverse axis** of the hyperbola. The y-axis does not intersect the hyperbola and is called its **conjugate axis.** The constants a and b are called the **semi-axes** of the hyperbola.

Besides, equation (27) shows that $\frac{x^2}{a^2} \geqslant 1$, that is x must be either $\leqslant -a$ or $\geqslant a$ (see Fig. 84).

A hyperbola has two asymptotes. We shall demonstrate this property restricting ourselves only to the case of the first quadrant (which is sufficient because of the symmetry properties of a hyperbola). From (27) it follows that

$$y_{hyp} = b \sqrt{\frac{x^2}{a^2} - 1} = \frac{b}{a} \sqrt{x^2 - a^2}$$

Later on (in particular, in the end of Sec. IV.22) we shall present some general rules for investigating expressions of this kind. But
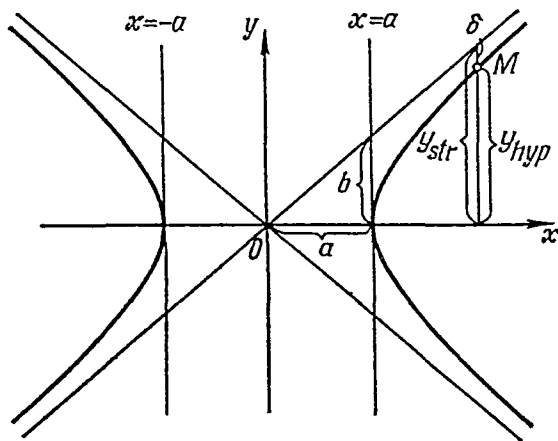
Fig. 84

we are not familiar with these rules yet and therefore let us apply certain artificial transformations which make it possible to "educe" $\frac{b}{a} x$ from $\sqrt{x^2 - a^2}$:

$$\frac{b}{a} \sqrt{x^2 - a^2} = \frac{b}{a} [x + (\sqrt{x^2 - a^2} - x)] = \frac{b}{a} x + \frac{b}{a} (\sqrt{x^2 - a^2} - x)$$

To investigate the behaviour of the second summand $\frac{b}{a} (\sqrt{x^2 - a^2} - x)$ as x increases let us multiply and, simultaneously, divide the summand by $\sqrt{x^2 - a^2} + x$. This yields

$$y_{hyp} = \frac{b}{a} x + \frac{b}{a} \frac{(\sqrt{x^2 - a^2} - x)(\sqrt{x^2 - a^2} + x)}{\sqrt{x^2 - a^2} + x} = \frac{b}{a} x - \frac{ab}{\sqrt{x^2 - a^2} + x}$$

The fraction entering as the second summand into the right-hand side unlimitedly approaches zero when the variable point M travels into infinity along the hyperbola. Now taking the straight

line $y_{str} = \dfrac{b}{a}\, x$ we see that the difference $\delta = y_{str} - y_{hyp}$ unlimitedly approaches zero as $x \to \infty$ and therefore this straight line is an asymptote of the hyperbola. Taking into account the symmetry we finally obtain the equations of the asymptotes:

$$y = \pm \frac{b}{a}\, x$$

It can be shown that the section of the surface of a right circular cone (infinitely prolonged in both directions) by a plane which forms an angle with the axis of the cone is a hyperbola provided the angle is less than the one formed by the axis and the slant side of the cone (see Fig. 85). Try to prove this assertion reasoning as in the end of Sec. 10.
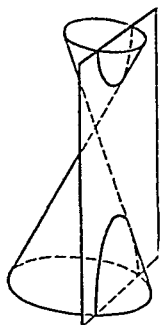
The properties of a hyperbola can be illustrated by the following example. Suppose a sound signal is issued from a point A. Let this signal be received at two points B and C. Suppose the signal is received τ sec earlier at B than at C. Then we can guarantee that the point A lies on a part of a hyperbola (the nearest to the point B) having its focuses at B and C and the transverse semi-axis $\dfrac{v_s\tau}{2}$ where v_s is the speed of sound (let the

Fig. 85

reader explain this fact). If two experiments of this kind are carried out the position of the point A is determined as the point of intersection of the corresponding hyperbolas.

12. **Relationship Between Ellipse, Hyperbola and Parabola.** There is a close relationship between an ellipse (see Sec. 10), a hyperbola (see Sec. 11) and a parabola (see Sec. I.23). This can be accounted for by the fact that all these curves are algebraic curves of the second order and, as it will be shown in Sec. 13, there are no other curves of the second order except for some singular cases similar to those discussed in Sec. 8. There are many problems involving parameters whose solution is one of these curves depending on the values of the parameters. In such circumstances the parabola (or its degenerated forms) usually occupies an intermediate position between the ellipse and the hyperbola.

Let us consider the intersection of a right circular cone (depicted in Fig. 86) with a plane which turns about an axis drawn perpendicularly to the axis of the cone (for example, about the axis pp). When the slope is slight (we mean the angle between this plane and the plane perpendicular to the axis of the cone) we have an elliptic section. The ellipse elongates as the slope increases and its

eccentricity also increases. When the plane intersects both parts of the double cone we have a hyperbolic section. Besides, we see that in the intermediate position when the plane is parallel to the slant side of the cone we have an intersection line which is infinite but still consists of one component. There will be no singular cases here similar to those indicated in Sec. 8 and, besides, a degeneration of an algebraic curve of the second order cannot yield a curve of a higher order. Hence the intersection line is a parabola. On the basis of these properties an ellipse, a hyperbola and a parabola are called conic sections.

Let us now apply the same point of view to the discussion of the simplest equation of an algebraic curve of the second order in polar coordinates. We begin with the polar equation of an ellipse. Let us put the pole O at the right focus (see Fig. 87). Applying the cosine law to the triangle AMO where A is the left focus and M is a variable point of the ellipse we obtain

$$AM^2 = AO^2 + OM^2 - 2AO \times$$
$$\times OM \cos (180° - \varphi);$$
$$(2a - \rho)^2 = (2c)^2 + \rho^2 +$$
$$+ 2 \cdot 2c\rho \cdot \cos \varphi;$$
$$4a^2 - 4a\rho + \rho^2 = 4(a^2 - b^2) +$$
$$+ \rho^2 + 4c\rho \cdot \cos \varphi$$

which implies

$$\rho = \frac{b^2}{a + c \cos \varphi} = \frac{\dfrac{b^2}{a}}{1 + \left(\dfrac{c}{a}\right) \cos \varphi}. \quad (28)$$

Denoting $\dfrac{b^2}{a} = p$ for brevity (p is called the **focal parameter of the ellipse**) we finally deduce the equation

$$\rho = \frac{p}{1 + \varepsilon \cos \varphi} \qquad (29)$$

(In case the pole is placed at the left focus we shall have the expression $1 - \varepsilon \cos \varphi$ in the denominator.) If now we take a hyperbola and place the pole at its left focus (see Fig. 88) then after similar transformations (which we leave to the reader) we arrive at the same formula (28). If we then denote $\dfrac{b^2}{a} = p$ and $\dfrac{c}{a} = \varepsilon$ the equation (29) will be obtained again. But in this case the dimensionless quantity ε (also called the **eccentricity of the hyperbola**) should be > 1.
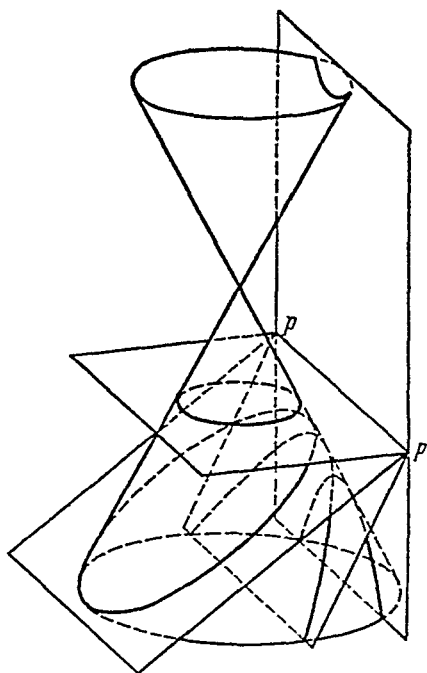
Fig. 86

It is easy to verify that if we take equation (29) with $\varepsilon = 1$ and pass from the polar coordinates to the Cartesian coordinates according to formulas (5) we obtain the equation of a parabola. In fact, $\rho = \dfrac{p}{1+\cos\varphi}$ and $\rho + \rho\cos\varphi = p$ which implies $\sqrt{x^2+y^2} + x = p$, $\sqrt{x^2+y^2} = p - x$ and $x^2 + y^2 = p^2 - 2px + x^2$. Therefore, finally,

$$x = -\frac{1}{2p}\, y^2 + \frac{p}{2}$$

(see Sec. I.23). Thus, equation (29) represents a parabola for $\varepsilon = 1$.
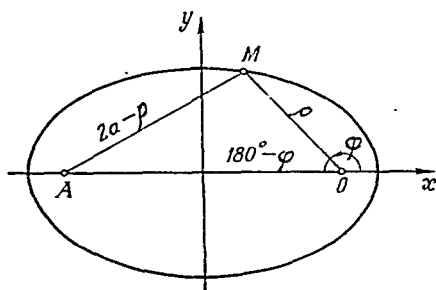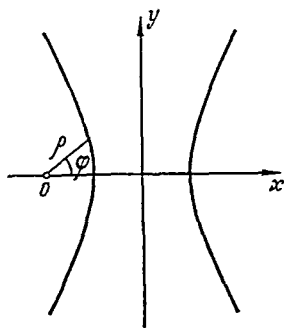
Fig. 87   Fig. 88

The pole of the polar coordinates which was introduced above is called the focus of the parabola. We see that a parabola, in contrast to an ellipse or a hyperbola, has only one focus. Now recall an interesting fact shown in Fig. 79: in case $\varepsilon = 1$ an ellipse turns into a line segment. Thus, the degeneration may yield different results in different problems.

Equation (29) is applied, in particular, to the problem of motion of two bodies under their mutual Newtonian attraction which is known as the *problem of two bodies* in celestial mechanics. Let us consider, for example, launching an artificial satellite of the earth from a point T (lying outside the earth's atmosphere) in the horizontal direction (see Fig. 89). If the initial velocity v_0 is not sufficient the satellite will not rotate round the earth. When the "first cosmic velocity" is achieved the satellite will rotate round the earth in a circular orbit with centre at the centre of the earth. If the velocity v_0 is then increased the rotation will be in an elliptic orbit and the centre of the earth will be at one of the focuses of the ellipse.

The further increase of the velocity v_0 makes the eccentricity of the ellipse increase and the second focus of the ellipse moves off

the first one. After the "second cosmic velocity" (the escape velocity) has been achieved the trajectory becomes parabolic and the satellite will not return to the point T. Therefore a parabola may be regarded as an ellipse with one of its focuses removed into infinity. The further increase of the velocity makes the trajectory turn into a hyperbola and thus the second focus appears again but this time "on the other side". The centre of the earth remains at one of the focuses of the orbit all the time.
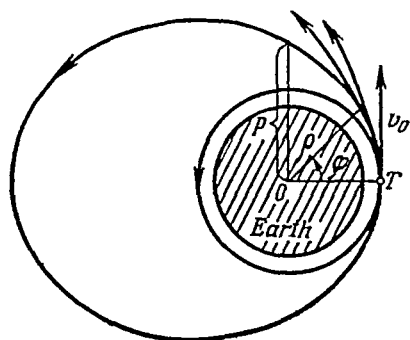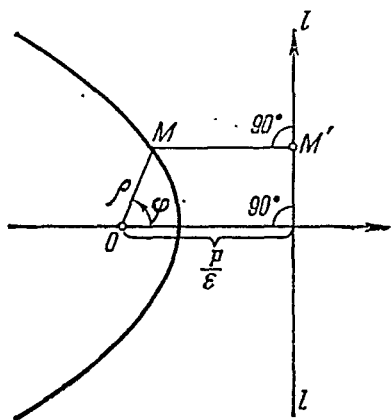
Fig. 89   Fig. 90

There is one more way of defining conic sections. Rewrite equation (29) in the form

$$\rho + \rho\varepsilon\cos\varphi = p \text{ or } \rho = p - \rho\varepsilon\cos\varphi = \varepsilon\left(\frac{p}{\varepsilon} - \rho\cos\varphi\right)$$

Now we see that the expression in the parentheses obtained above is just equal to the length of the line segment MM' shown in Fig. 90 (verify this!) where the straight line ll is drawn perpendicularly to the polar axis at the distance $\frac{p}{\varepsilon}$ from the pole. But $\rho = OM$, that is we obtain $OM = \varepsilon MM'$ which implies $\frac{MO}{MM'} = \varepsilon = \text{const.}$ Thus, an ellipse, a hyperbola and a parabola can be defined in a new way as *the locus of all points in a plane for which the ratio of their distances from a certain point (which is a focus) to their distances from a certain straight line* (the so-called **directrix**) *is a constant quantity.*

13. General Equation of a Curve of the Second Order. Let us put down the general form of an equation of an algebraic curve of the second order [as an analogue to equation (18)]:

$$Ax^2 + 2Bxy + Cy^2 + Dx + Ey + F = 0 \qquad (30)$$

(we write $2B$ instead of B because, as it will be seen, this simplifies some formulas which will be further obtained). Now our aim is to transform the Cartesian coordinates (see Sec. 2) in such a way that equation (30) should take the simplest form; this will enable us to find out what curve is determined by the equation.

The canonical equation containing no terms with the product of the coordinates, we first of all try to turn the coordinate axes in such a manner that this product should be eliminated. According to formulas (4), after the axes are turned through an angle α, the equation in the new coordinates will have the form

$$A \, (x' \cos \alpha - y' \sin \alpha)^2 + 2B \, (x' \cos \alpha - y' \sin \alpha) \, (x' \sin \alpha +$$
$$+ \, y' \cos \alpha) + C \, (x' \sin \alpha + y' \cos \alpha)^2 + D \, (x' \cos \alpha - y' \sin \alpha) +$$
$$+ \, E \, (x' \sin \alpha + y' \cos \alpha) + F = 0 \qquad (31)$$

Removing the parentheses and collecting the terms containing the product $x'y'$ we see that the coefficient in $x'y'$ is equal to

$$-2A \cos \alpha \sin \alpha + 2B \cos^2 \alpha - 2B \sin^2 \alpha + 2C \sin \alpha \, \cos \alpha =$$
$$= 2B \cos 2\alpha + (C - A) \sin 2\alpha$$

Thus, we must have

$$2B \cos 2\alpha + (C - A) \sin 2\alpha = 0, \quad \text{i.e.} \quad \tan 2\alpha = \frac{2B}{A - C} \qquad (32)$$

Now we find the angle α from equation (32) and thus determine through what angle the coordinate axes should be turned.

After the axes are turned through the angle α the equation takes the form

$$A'x'^2 + C'y'^2 + D'x' + E'y' + F = 0 \qquad (33)$$

where A', C', D' and E' are some coefficients which can be found by collecting similar terms in (31). But, as it will be proved in Sec. XI.11, a rotation of the coordinate axes does not change the quantity $AC - B^2$ although the coefficients A, B and C in terms of the second degree may vary, that is the expression is an invariant. There is no term with $x'y'$ in equation (33) and therefore, since $B' = 0$, we obtain

$$A'C' - B'^2 = A'C' = AC - B^2$$

Consequently, if the expression $AC - B^2$ written for original equation (30) is positive then the coefficients A' and C' in equation (33) have the same signs since their product is positive. Such a case is called **elliptic**. If $AC - B^2 < 0$ then A' and C' are of opposite signs (this is a **hyperbolic** case). Finally, if $AC - B^2 = 0$ then one of the coefficients (A' or C') is equal to zero (the so-called **parabolic** case).

Let us turn to the elliptic case. Completing the square in equation (33) (compare with Sec. 4) we arrive at an equation of the form

$$A' (x' - a)^2 + C' (y' - b)^2 + F' = 0$$

Let us now translate the axes x' and y' (see Sec. 2) and pass to the new coordinates $x'' = x' - a$ and $y'' = y' - b$. Then the equation turns into

$$A'x''^2 + C'y''^2 = -F', \quad \text{that is} \quad \frac{x''^2}{-\dfrac{F'}{A'}} + \frac{y''^2}{-\dfrac{F'}{C'}} = 1 \quad (34)$$

For the sake of definiteness let us suppose that both A' and C' are positive. Then if $F' < 0$ we obtain the canonical equation of an ellipse. Consequently, the original curve is an ellipse but displaced and turned with respect to the axes x and y. If $F' > 0$ or $F' = 0$ we obtain the singular cases mentioned in Sec. 8 since there will be, respectively, either an imaginary curve or a single point.

Similarly, in the hyperbolic case the curve must be a hyperbola and in the parabolic case the curve must be a parabola with the exception of the singular cases described in Sec. 8 which, as a rule, are of no significance.
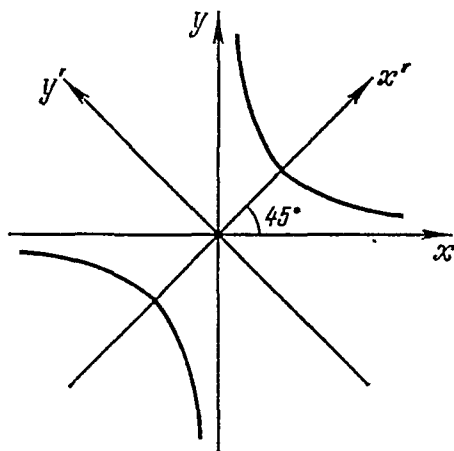
Fig. 91

As an example, let as again consider the graph of the function $y = \dfrac{k}{x}$ which expresses inverse variation (see Sec. I.25). The equation can be rewritten as $xy - k = 0$. Comparing with equation (30) we see that in this case $A = C = D = E = 0$, $B = \dfrac{1}{2}$ and $F = -k$. Since $AC - B^2 = -\dfrac{1}{4} < 0$ we have a hyperbolic case. Formula (32) now implies $\tan 2\alpha = \dfrac{1}{0} = \pm\infty$ and this means that we can put $2\alpha = \dfrac{\pi}{2}$ or $\alpha = \dfrac{\pi}{4}$. Let us rotate the axes through the angle of 45°. According to formulas (4) we have

$$x = x' \frac{\sqrt{2}}{2} - y' \frac{\sqrt{2}}{2} = \frac{\sqrt{2}}{2} (x' - y'),$$

$$y = x' \frac{\sqrt{2}}{2} + y' \frac{\sqrt{2}}{2} = \frac{\sqrt{2}}{2} (x' + y')$$

Substituting the expressions into the original equation we obtai

$$\frac{\sqrt{2}}{2}(x'-y')\frac{\sqrt{2}}{2}(x'+y')-k=0$$

i.e.

$$\frac{x'^2-y'^2}{2}-k=0, \quad \frac{x'^2}{2k}-\frac{y'^2}{2k}=1$$

Consequently, the curve in question is a hyperbola which has equa semi-axes: $a = b = \sqrt{2k}$ (see Fig. 91). The fact we have establishe here accounts for the usage of the term "hyperbola" in Sec. I.25

Limit. Continuity

§ 1. Infinitesimal and Infinitely Large Variables

1. Infinitesimal Variables. Infinitesimal variables form a very important class of variables which is of great significance in higher mathematics. *A variable changing in a certain process is called* infinitesimal *if in this process it approaches (tends to) zero unlimitedly.* For instance, let us consider the process of expansion of a given mass of gas. If the gas expands unlimitedly then its density and pressure are infinitesimals; this is an example of positive, continuous and monotonic (see Sec. I.5) infinitesimal variables. In the process of damped oscillations of a pendulum its angle of deviation from the equilibrium position also represents an infinitesimal variable as time increases but this variable is an oscillating one and assumes both positive and negative values (and the zero value as well). If we take the sequence $a_1 = -\frac{1}{1^2}$, $a_2 = -\frac{1}{2^2}$, $a_3 = -\frac{1}{3^2}$, ... then its general term $a_n = -\frac{1}{n^2}$ is a discrete and negative infinitesimal variable in the process in which the number n increases: $n = 1, 2, 3, \ldots$. It should be noted that when a variable is qualified as an infinitesimal one we must point out a certain process in which the variable changes since the same variable may not be an infinitesimal at all in some other process.

As it has been stated an infinitesimal variable α "approaches zero unlimitedly". Let us discuss this term in detail. Suppose a variable α changes in some process and tends to zero. Then there must exist a moment from which on we shall necessarily have $|\alpha| < 1$; similarly, there must exist some other (later) moment from which on $|\alpha| < 0.1$. By the same reasoning, there must exist the third moment (following the first two moments) from which on we shall always have $|\alpha| < 0.01$ and so on. This can be expressed in the following manner: for any $\varepsilon > 0$ there must exist a certain moment in the development of the process from which on there will always be $|\alpha| < \varepsilon$. It is not necessary to indicate such a mo-

ment virtually in all cases: we should only be sure of the very existence of the moment. Hence, an infinitesimal variable may not be small at all at the beginning of the process in which it changes and the essential fact is that it becomes arbitrarily small (in its absolute value) when the process goes on sufficiently long.

In addition, let us now state more precisely what the expression "a moment" in the development of a process means. If we consider a process developing in time then "a moment" simply means a certain moment of time. The realization of a process may not be connected with the time but may be related to some other variable (for instance, in the third of the above examples the process is characterized by the change of the number n which assumes the successive values 1, 2, 3, . . .); in such cases the existence of "a moment" means that a variable which characterizes the process takes on a certain value.

Taking into account the specifications given in the last two paragraphs we can say that the general term a_n of a sequence is an infinitesimal variable in the process which is characterized by the increase of the number n if for any arbitrarily chosen $\varepsilon > 0$ it is possible to indicate a number $N = N(\varepsilon)$ such that there will be $|a_n| < \varepsilon$ as n becomes larger than N ($n > N$).

The notion of an infinitesimal variable can be specified in an analogous manner for other types of variables and processes but we shall not use them further.

From the point of view of the definition a constant value, even a very small one, is not an infinitesimal variable; only the constant value equal to zero is an infinitesimal variable from the formal point of view of the above definition.

Now we must point out that the definition of an infinitesimal variable which we shall use here involves the following principal difficulty: there are no real variables which may approach zero unlimitedly. Indeed, in the examples considered above the gas cannot expand unlimitedly and the real pendulum stops oscillating after some time has passed. Furthermore, if we take into account the molecular structure of a substance we see that we cannot take an infinitely small mass of the substance because it is impossible to consider the mass of the substance which is smaller than the mass of its molecule; similar situations are observed in other examples.

Thus, our definition applies only to a mathematical model of a real process in which the real situation is simplified so that the application should become possible. Therefore we speak about a pendulum which has oscillations lasting infinitely long and about a "continuous" (non-molecular) structure of a substance and so forth. We always replace a real process by its mathematical model but this must be performed in such a way that the main features of the process we are interested in should not be considerably changed.

But nevertheless we always deal with a model and we must not forget it. Otherwise some principal mistakes may occur. For example, one can make an attempt to attribute all the properties of a model to the reality without sufficient reasons.

There is another possibility of interpreting the usage of the notion of an infinitesimal in practical applications. Let us discuss it here. The practical ("physical") infinity should be distinguished from the mathematical concept of infinity. Thus, a "practical" infinitesimal is a variable or even a constant which is sufficiently small in comparison with "finite" values which are involved in a certain investigation, i.e. so small that it should be possible to apply to it all the properties of "mathematical" infinitesimals without any considerable error. At the same time this value must not be too small, that is so small that it should be necessary to take into account some effects of the microstructure when it is inexpedient, or so small that it should not comply with the real possible values. For example, suppose we are studying the deformations of an elastic body; then certain sizes which are sufficiently small in comparison with the size of the body and, at the same time, sufficiently large in comparison with the molecular sizes may be regarded as infinitesimals etc.

In what follows we shall use the definition given in the beginning of this section but from time to time we shall come back to the considerations discussed here.

2. Properties of Infinitesimals. The properties of infinitesimals are directly implied by the definition given in Sec. 1.

1. *The sum or the difference of two infinitesimals is also an infinitesimal variable.* Indeed, if each summand approaches zero then the sum does the same. Similarly, the sum of three, ten and, in general, of an arbitrary finite number of infinitesimals is also an infinitesimal. We point out here that there are some circumstances when in the development of a process the number of summands entering into the sum increases infinitely; then even if each summand is an infinitesimal variable the whole sum may not be infinitesimal. For instance,

$$\underbrace{\frac{1}{n^2}+\frac{1}{n^2}+ \cdots + \frac{1}{n^2}}_{n \text{ times}}=\frac{1}{n}$$

$$\underbrace{\frac{3}{n}+\frac{3}{n}+ \cdots + \frac{3}{n}}_{n \text{ times}}=3$$

$$\underbrace{\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{n}}+ \cdots + \frac{1}{\sqrt{n}}}_{n \text{ times}}=\sqrt{n}$$

Here we have the situation described above when n increases; but the first sum is an infinitesimal variable, the second sum is constant and the third sum even increases unlimitedly.

2. *The product of an infinitesimal variable by a bounded variable* (see Sec. I.5) *is again an infinitesimal variable.* Let, for example, the first factor be always in the limits from 0 to 1000 and let the second factor assume the successive values 1, 0.1, 0.01, 0.001 and so on. Then the values of the product will be, respectively. smaller than $1000 \times 1 = 1000$, 100, 10, 1, 0.1, 0.01, 0.001 and so forth.

In particular, this property implies that *the product of an infinitesimal by a constant is an infinitesimal variable. The product of two infinitesimals is an infinitesimal* since an infinitesimal variable is, of course, a special case of a bounded variable. Similarly, the product of any arbitrary number of infinitesimals is an infinitesimal variable.

We remark that *the ratio of two infinitesimals may not be an infinitesimal.* If, for example, $\alpha = \frac{1}{n}$, $\beta = \frac{1}{n^2}$ and $\gamma = \frac{1}{n} + \frac{1}{n^2}$ where n takes successive values 1, 2, 3, ... then the variables α, β and γ are infinitesimals. But at the same time the first of the ratios $\frac{\beta}{\alpha} = \frac{1}{n}$, $\frac{\gamma}{\alpha} = 1 + \frac{1}{n}$ and $\frac{\alpha}{\beta} = n$ is an infinitesimal while the second approaches 1 and the third even increases unlimitedly. We shall discuss such ratios in detail in § 3.

3. **Infinitely Large Variables.** *A variable x is called* infinitely large *in some process of changing if it increases in this process in its absolute value unlimitedly;* then we write $|x| \to \infty$. An infinitely large variable may be positive and then we write $x \to +\infty$ or negative (then we write $x \to -\infty$) but it may also change its sign: for instance, the variable $x_n = (-2)^n$ assumes the values -2, 4, -8, 16. ... as the number n increases and therefore it is infinitely large but we cannot say that $x_n \to \infty$ or $x_n \to -\infty$. The comprehensive statement of the notion "increases infinitely" is analogous to the one given in Sec. 1 for the notion of "infinitely approaches zero" but, of course, here we should consider inequalities of the form $|x| > N$. This means that from a certain moment on the variable must satisfy the inequality $|x| > 1$ and from some other (later) moment on it must satisfy the inequality $|x| > 10$. Further. there must exist a certain moment from which on we shall have $|x| > 100$ and so on.

Now let us discuss some simple properties of infinitely large variables. *A variable which is the inverse of an infinitely large variable is an infinitesimal and, conversely, the inverse of an infinitesimal variable is an infinitely large variable.* These properties can be conditionally expressed as

$$\frac{1}{\infty} = 0 \quad \text{and} \quad \frac{1}{0} = \pm \infty$$

We shall use this notation but one must understand it correctly. For example, the first of the properties means that if the variable x entering into the equality $\frac{1}{x} = \alpha$ increases unlimitedly then in the same process the variable α approaches zero (or, as in Sec. 1, if x is a "practical" infinitely large variable then α is "practically" infinitesimal). All formulas containing the symbol of infinity ∞ should be understood in a similar way. For instance, the formula $\tan \frac{\pi}{2} = \pm \infty$ is a conditional and abbreviated form of expressing the fact that when the variable φ approaches $\frac{\pi}{2}$ in some process then the variable $x = \tan \varphi$ increases unlimitedly in its absolute value, that is x is infinitely large and so on. This enables us to operate on the symbol ∞ as if it were a usual number in many cases but, of course, ∞ is not a concrete number but only a symbol indicating infinitely large variables which are different in different circumstances.

The sum of an infinitely large variable and a bounded variable is infinitely large since the first summand "gains over". The sum of two infinitely large variables of a similar sign is also infinitely large. In contradistinction to it the sum of two infinitely large variables of opposite signs may not be infinitely large since these infinitely large variables may "compensate" mutually. These facts are written as $\infty + \infty = \infty$; the expression $\infty - \infty$ denotes an **indeterminate form**. This shows that it is impossible to operate on the symbol ∞ as on a usual number in all cases; not always $\infty - \infty = 0$ since $\infty - \infty$ is an abbreviated and conditional way of denoting a difference of the form $X - Y$ where X and Y are infinitely large variables. The behaviour of these infinities may vary in different cases and therefore it is impossible to have a good judgment on the behaviour of their difference unless an additional investigation has been carried out. We shall discuss indeterminate forms of various types in detail in our course later on.

The product of two infinitely large variables is an infinitely large variable. Moreover, the product of an infinitely large variable by a variable which is larger than a positive constant in its absolute value is an infinitely large variable. At the same time, the ratio of two infinitely large quantities is an indeterminate form like the ratio of two infinitesimals.

§ 2. Limits

4. Definition. It is said that *a variable x approaches (tends to) a finite limit a in some process if a is constant and x approaches a unlimitedly in this process.* Then we write

$$x \to a \quad \text{or} \quad \lim x = a$$

Thus, the limit of a variable in case it exists is a constant value.

According to the above definition infinitesimals are variables that approach zero, that is having zero as a limit. But infinitely large variables, of course, have no finite limit.

To say "x approaches a unlimitedly" is to say "the difference between x and a approaches zero unlimitedly", that is $x - a = \alpha$ is an infinitesimal. The last equality may be rewritten as

$$x = a + \alpha$$

where α is the infinitesimal.

If a variable x approaches its limit a but always remains smaller than a, that is approaches a from the region of smaller values, then we conditionally write $x \to a - 0$ or $\lim x = a - 0$ (this is a conditional way of denoting the limit since if we understand the
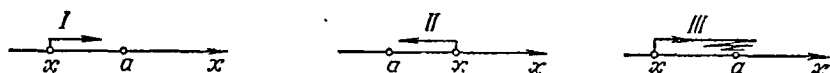
Fig. 92

expression $a - 0$ as a real difference then $a - 0 = a$). If x in its process of approaching a always remains larger than a then we write $x \to a + 0$. Finally, x may tend to a in such a way that it could take on values larger than a and values smaller than a all the time (such a process resembles damped oscillations). All the cases described here are depicted in Fig. 92.
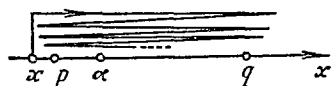
Fig. 93

Now we can sum up our discussion on the types of variables. A variable x may be of one of the following types in a certain process:

(1) x is bounded and has a limit; in a special case when the limit is equal to zero x is an infinitesimal variable. To distinguish between these cases a bounded variable is sometimes called finite only if it is not an infinitesimal; for instance, it is possible to speak about an infinitesimal mass and about a finite mass etc.

(2) x is bounded but has no limit; as an example we may consider the deviation of a pendulum from its equilibrium position in the case of undamped oscillations. Variables of this type are called **oscillating** (see Fig. 93).

In Fig. 93 we see the point α which possesses the following property: the variable x approaches the point α infinitely many times in the process of its change and takes on the values that are arbitrarily close to α but at the same time x does not remain near α all the time. In this case α is called a **limit point** of the variable x.

There are the greatest value q and the least value p among these limit points. q and p are denoted, respectively, as $\overline{\lim}\, x$ and $\underline{\lim}\, x$ and are called the **limit superior** and the **limit inferior** of the variable x. But in this case x does not have a "unique" limit which we discussed in the beginning of this section. Therefore we should remark that the everyday notion of a "limit" (in the sense of some kind of a "border", "frontier") differs from the mathematical one.

A bounded variable x always has the limit superior and the limit inferior and $\underline{\lim}\, x \leqslant \overline{\lim}\, x$. The "unique" limit, that is the limit in the sense of our previous definition, exists if and only if $\underline{\lim}\, x =$
$= \overline{\lim}\, x$.

(3) x is unbounded and besides infinitely large. In this case we write $\lim x = \pm\infty$, and x is said to have an **infinite limit**.
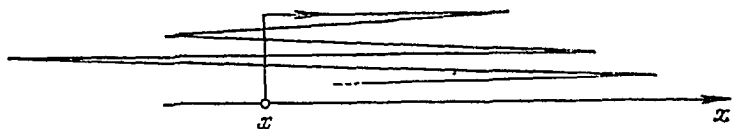
Fig. 94

(4) x is unbounded but not infinitely large. The deviation of an oscillating body in the case of a resonance may serve as an example here. Such a variable oscillates and from time to time travels "toward infinity" further and further but at the same time it permanently returns to regions lying near the initial point (see Fig. 94).

5. Properties of Limits.

1. *If a variable has a limit then this limit is unique*, i.e. there exist no other limits of the variable (the property is obviously implied by the definition).

2. *The limit of a constant equals the constant* (the property is obvious).

3. If $x \to a$ and $y \to b$ in one and the same process then $x + y \to$ $\to a + b$. This can be written in a different manner as

$$\lim (x + y) = \lim x + \lim y \qquad (1)$$

and formulated as follows: *the limit of a sum is equal to the sum of the limits*. To prove this let us write $x = a + \alpha$ and $y = b + \beta$ where α and β are infinitesimals. Then $x + y = (a + b) + (\alpha + \beta)$. Hence, the variable $x + y$ is represented as a sum of the constant $a + b$ and the infinitesimal $\alpha + \beta$ (see Sec. 2). Therefore, $(x + y) \to$ $\to (a + b)$.

The result thus obtained may be interpreted "practically" as the following example: if $3.002 \approx 3$ and $2.001 \approx 2$ then $5.003 \approx 5$

4. The limit of a product equals the product of the limits. A more complete statement is the following: *if the factors entering in a product have limits then the whole product also has a limit which is equal to the product of the limits of the factors.* Indeed, using our previous notation we have $xy = ab + (a\beta + b\alpha + \alpha\beta) \to ab$, that is

$$\lim (xy) = \lim x \lim y \qquad (2)$$

Here, as in property 3, we have taken only two variables but it is easy to verify that these properties remain true for any arbitrary finite and constant number of variables. For example,
$$\lim (xyz) = \lim [(xy) z] = \lim (xy) \lim z = \lim x \lim y \lim z$$

5. A constant factor may be taken outside the sign of the limit, that is $\lim (Cx) = C \lim x$ where $C = \text{const}$. This property follows from properties 2 and 4.

6. The limit of a ratio is equal to the ratio of the limits, i.e.

$$\lim \left(\frac{x}{y} \right) = \frac{\lim x}{\lim y} \qquad (3)$$

with the exception of those cases when both the numerator and the denominator tend to zero, that is when we have the indeterminate form $\frac{0}{0}$.

To prove this we first suppose that $\lim y = b \neq 0$. Then

$$\frac{x}{y} = \frac{a + \alpha}{b + \beta} = \frac{a}{b} + \left(\frac{a + \alpha}{b + \beta} - \frac{a}{b} \right) = \frac{a}{b} + \frac{\alpha b - \beta a}{b (b + \beta)}$$

The numerator of the last fraction is infinitesimal whereas the denominator $\approx b^2 = \text{const} \neq 0$ and therefore the whole fraction is infinitesimal while the first fraction is constant. This implies our assertion.

In case $\lim y = 0$ and $\lim x \neq 0$ we have $\frac{x}{y} = \frac{1}{y} x \to \pm\infty$ (see Sec. 3). Therefore we obtain $\pm\infty$ on either side of formula (3).

By Sec. 3, formulas (1), (2) and (3) hold not only for finite limits but also for infinite ones with the exception of those cases when there are indeterminate forms of types $\infty - \infty$, $0 \cdot \infty$ and $\frac{\infty}{\infty}$ on the right-hand sides. These forms will be discussed in Secs. III.3 and IV.4.

7. If $x \to a$ and $a > 0$ then x becomes and remains greater than zero $(x > 0)$ as the process of its change lasts sufficiently long (i.e. from some moment on). This obviously follows from the definition of a limit.

8. It is permissible to pass to the limit in an inequality: if $x \leqslant y$ then $\lim x \leqslant \lim y$ (naturally, if these limits exist). Indeed, let us denote $z = y - x$. Then $z \geqslant 0$ and therefore z cannot approach

a constant which is negative. Hence, $\lim z \gg 0$, $\lim (y - x) \gg 0$ and $\lim y - \lim x \gg 0$.

We remark here that if $x < y$ then after the passage to the limit we can obtain either $\lim x < \lim y$ or $\lim x = \lim y$ because the difference between x and y may tend to zero. Thus, we cannot retain the strict inequality unless an additional investigation has been carried out.

9. If $x \leqslant y \leqslant z$ and we have $x \to a$ and $z \to a$ in one and the same process then $y \to a$ (see Fig. 95).

10. If a variable x increases monotonically then it either increases unlimitedly, i.e. $x \to +\infty$, or is bounded and then has a finite limit: $x \to a - 0 < \infty$. If, in addition, x has an upper bound A
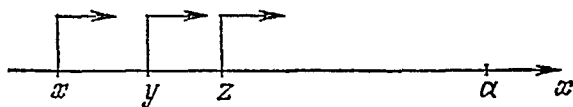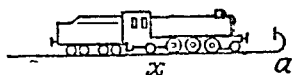
Fig. 95   Fig. 96

(i.e. $x \leqslant A$) then $\lim x = a \leqslant A$. A monotonically decreasing variable behaves similarly. These obvious assertions (see Fig. 96) are indeed the expression of an essential property of the "completeness" of the totality of all real numbers. If we used only rational numbers all the preceding properties of limits would remain true with the exception of property 10 since rational values may lead to an irrational result when passing to the limit. The rigorous justification of property 10 may be found, for example, in [14].

Thus, *a bounded monotonic variable must have a finite limit*; a bounded but non-monotonic variable may have no limit (see Sec. 4).

To conclude the section we point out that the dimension of a variable quantity remains the same when we pass to the limit: if $x \to a$ then $[x] = [a]$.

The first attempt to create the theory of limits was made by Newton in 1686 but in fact the operation of passing to the limit had been used earlier beginning with Greek scientists. The notion of a limit which is close to the one used in this book was formulated in 1765 by J. D'Alembert (1717-1783), a French mathematician, philosopher and enlightener of the pre-revolutionary period in France.

6. Sum of a Numerical Series. The idea of a limit is directly applicable to an important notion of a sum of a series. As a preliminary let us introduce the following abbreviated notation:

$$a_p + a_{p+1} + a_{p+2} + \ldots + a_{q-1} + a_q = \sum_{k=p}^{q} a_k \qquad (4)$$

Here $\sum\limits_{k=p}^{q}$ is the **summation sign** which indicates that we should substitute $k = p$, $p + 1$, ..., q into the expression following the sign and then sum up the results ($\sum$ is the Greek letter *sigma*), a_k is the general term, or general element, or kth term of the sum (of the series), k is the number of the term (the **index of summation**), p and q are, respectively, the **lower** and the **upper limits of summation** showing the range of the index k. For example,

$$\sum_{k=3}^{8} \frac{1}{k^2} = \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2} + \frac{1}{8^2}\ (= 0.2774)$$

We should point out immediately that the sum does not depend on the notation of the index of summation, that is

$$\sum_{k=p}^{q} a_k = \sum_{i=p}^{q} a_i = \sum_{j=p}^{q} a_j = \ldots$$

Virtually, all these sums are equal to the left-hand side of (4). Thus, the summation index is a **dummy index**, that is it does not enter in the result and may be denoted by any letter.

Now we turn to "infinite sums" or, more precisely, to the notion of a numerical series. A numerical series is an infinite expression of the form

$$a_1 + a_2 + \ldots + a_n + \ldots = \sum_{k=1}^{\infty} a_k \qquad (5)$$

and the summands a_1, a_2, a_3, ... are certain numbers called the **terms of the series**. To define the sum of series (5) it is necessary first to compose the so-called "partial sums" of series (5):

$$S_1 = a_1; \quad S_2 = a_1 + a_2; \quad S_3 = a_1 + a_2 + a_3; \quad \ldots; \quad S_n = \sum_{k=1}^{n} a_k; \quad \ldots$$

If the nth partial sum tends to a certain finite limit as the number n increases then series (5) is called **convergent** and its sum S is understood as

$$S = \sum_{k=1}^{\infty} a_k = \lim_{n \to \infty} S_n = \lim_{n \to \infty} \sum_{k=1}^{n} a_k$$

(The inscriptions in small type under the sign lim and, in other cases, under the sign $\to$ indicate the process in which the limits are considered.) Partial sums of a convergent series which have large numbers are practically equal to each other and to the whole sum of the series. If there exists no finite limit of partial sums series

(5) is called **divergent. In** particular, if partial sums approach infinity series (5) is said **to diverge to infinity;** in this case we write

$$\sum_{k=1}^{\infty} a_k = \infty \quad (\text{or } -\infty)$$

A divergent series has no finite sum.

The product of an infinite number of factors is determined in a similar way. The same manner of reasoning applies to any infinite process: first a finite process is performed and then the passage to the limit is carried out.

Let us consider, for example, the series

$$1 + \frac{1}{3} + \frac{1}{3^2} + \cdots + \frac{1}{3^n} + \cdots \tag{6}$$

Using the formula for the sum of a geometrical progression we obtain

$$S = \lim_{n \to \infty} S_n = \lim_{n \to \infty} \left(1 + \frac{1}{3} + \frac{1}{3^2} + \cdots + \frac{1}{3^{n-1}} \right) = \lim_{n \to \infty} \frac{1 - \frac{1}{3^n}}{1 - \frac{1}{3}} =$$

$$= \frac{1}{1 - \frac{1}{3}} = \frac{3}{2}$$

Thus, series (6) converges and its sum equals 1.5. If we calculate the partial sum of the first ten terms we receive approximately 1.499975.

In like manner the series

$$a + aq + aq^2 + \cdots + aq^n + \cdots \tag{7}$$

converges for $|q| < 1$ and its sum (the sum of an infinitely decreasing geometrical progression) is equal to $a(1-q)^{-1}$.

The nth partial sum S_n of the series

$$1 + 1 + 1 + \cdots + 1 + \cdots \tag{8}$$

equals n and therefore it tends to infinity. Hence, series (8) diverges to infinity. Similarly, $-1 - 1 - 1 - \cdots - 1 - \cdots = -\infty$.

Partial sums of the series

$$1 - 1 + 1 - \cdots + (-1)^{n+1} + \cdots \tag{9}$$

are equal, in succession, to $S_1 = 1$, $S_2 = 0$, $S_3 = 1$, $S_4 = 0$, $\ldots$ and they have neither a finite nor an infinite limit but remain bounded and oscillate without damping (compare with the end of Sec. 4, Case 2). Thus, series (9) diverges in an "oscillating" manner.

If we drop or add a term in a series (5) this will not affect the very fact of its convergence or divergence, that is if series (5) converged before, it will converge now though its sum may change and if series

(5) did not converge it will diverge after the operation. Indeed, if we take the series $a_2 + a_3 + \ldots + a_n + \ldots$ and consider it together with (5) then its partial sums will differ from the corresponding partial sums of (5) in the constant number a_1 and therefore if one of the sums approaches a limit the other does the same. Repeating such droppings or additions of a finite number of terms of series (5) we come to the conclusion that an arbitrary change of a finite number of terms of series (5) does not affect its convergence or divergence.

If series (5) converges then the series $R_n = a_{n+1} + a_{n+2} + a_{n+3} + \ldots$ also converges (why is it so?). Its sum is called the remainder (remainder term, remainder after n terms, or "tail") of series (5). It is clear that $S = \sum\limits_{k=1}^{\infty} a_k = \sum\limits_{k=1}^{n} a_k + \sum\limits_{k=n+1}^{\infty} a_k = S_n + R_n$.
It follows that the remainder of a convergent series tends to zero as the number n increases since it is the difference between the nth partial sum of the series and the limit of the partial sum.

We shall now establish *the necessary condition (test) for the convergence of series* (5). Since $S_{n-1} = a_1 + a_2 + \ldots + a_{n-1}$ and $S_n = a_1 + a_2 + \ldots + a_{n-1} + a_n$ we have $a_n = S_n - S_{n-1}$ and therefore

$$a_n \xrightarrow[n\to\infty]{} S - S = 0 \qquad (10)$$

if series (5) converges. Thus, *the general term a_n of series* (5) *tends to zero as the number increases*. This condition is not at all sufficient for the convergence; for example, the condition is fulfilled for the series

$$1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \ldots + \frac{1}{\sqrt{n}} + \ldots$$

but the series diverges to infinity. The divergence is implied by the fact that

$$S_n = 1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \ldots + \frac{1}{\sqrt{n}} > \underbrace{\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} + \ldots + \frac{1}{\sqrt{n}}}_{n \text{ times}} =$$

$$= n \cdot \frac{1}{\sqrt{n}} = \sqrt{n} \xrightarrow[n\to\infty]{} \infty$$

We shall systematically study series of different types in Chapter XVII where, in particular, some rigorous sufficient conditions (tests) for their convergence will be formulated and proved. Yet we are going to make use of some series before that as the question on their convergence may be settled in some practical sense, although not rigorously enough, in the following way. We compute the terms

one after another and when we see that soon enough, from some moment on, their values become less than the required degree of accuracy of calculations and that there is no reason to expect that the addition of the following terms may noticeably change the sum we simply drop all the subsequent terms. This means that we replace the series by a finite number of its terms (i.e. by a partial sum). In such a case we may say that the series is "practically convergent".

§ 3. Comparison of Variables

7. Comparison of Infinitesimals. Comparison of two infinitesimals with each other is carried out by investigating their ratio. Let $\alpha \neq 0$ and $\beta \neq 0$ be two infinitesimals varying in one and the same process. Then we can have the following cases.

(1) If $\dfrac{\beta}{\alpha} \to 0$ then β is said to tend to zero faster than α or that β is an **infinitesimal of higher order** than α and α is an **infinitesimal of lower order** than β. This fact is written in the form $|\beta| \ll |\alpha|$ or $|\alpha| \gg |\beta|$; the symbolic equality $\beta = o(\alpha)$ is also used for this purpose. Hence, in this case β is not only an infinitesimal but is also an infinitesimal part of the other infinitesimal α. For example, let ω be the volume of an infinitesimal cube and σ the volume of a right prism with the same base and with the constant altitude a. Then $|\omega| \ll |\sigma|$ since if h denotes the edge of the cube then

$$\omega = h^3, \quad \sigma = ah^2 \text{ and } \frac{\omega}{\sigma} = \frac{h^3}{ah^2} = \frac{h}{a} \to 0 \text{ as } h \to 0.$$

(2) If $\dfrac{\beta}{\alpha} \to \infty$ then $\dfrac{\alpha}{\beta} = \dfrac{1}{\beta/\alpha} \to 0$, i.e. $|\alpha| \ll |\beta|$.

(3) If the ratio $\dfrac{\beta}{\alpha}$ approaches a finite nonzero limit then α and β are called **infinitesimals of the same order**; in this case neither of the variables α and β can become much smaller than the other. In particular, in case this limit is equal to unity the variables α and β are called **equivalent infinitesimals**; then we write $\alpha \sim \beta$. For example, the infinitesimals x and $x + x^2$ are equivalent as $x \to 0$ whereas the infinitesimal variables $2x$ and $x + x^2$ are of the same order in this process but not equivalent since $\dfrac{2x}{x+x^2} \to 2$.

It is possible to verify the following properties.

(1) If α and β are of the same order and $|\gamma| \ll |\alpha|$ then $|\gamma| \ll |\beta|$.

(2) If α and β are of the same order and β and γ are also of the same order then the same is true for α and γ.

(3) If α and β are of the same order and have the same sign then $\alpha + \beta$ is of the same order as α and β; in case α and β have opposite signs $\alpha + \beta$ may happen to be of higher order and so forth.

For example, let us verify the first property:

$$\lim \frac{\gamma}{\beta} = \lim \left(\frac{\gamma}{\alpha} \frac{\alpha}{\beta} \right) = \lim \frac{\gamma}{\alpha} \lim \frac{\alpha}{\beta} = 0 \cdot \lim \frac{\alpha}{\beta} = 0$$

which implies what was required. (Verify the remaining properties; by the way, they are quite evident.)

(4) If $\frac{\beta}{\alpha} \to k \neq 0$ and we denote $\beta - k\alpha = \gamma$, i.e. $\beta = k\alpha + \gamma$, then $| \gamma | \ll | \alpha |$; in other words, infinitesimals of the same order are proportional to each other within the accuracy to a term of higher order. This follows immediately from

$$\frac{\gamma}{\alpha} = \frac{\beta - k\alpha}{\alpha} = \frac{\beta}{\alpha} - k \to 0$$

Such a case of "almost proportional" infinitesimals is sometimes denoted as $\alpha \sim \beta$.

8. Properties of Equivalent Infinitesimals. The following simple properties are true:

(1) if $\alpha \sim \beta$ then $\beta \sim \alpha$;

(2) if $\alpha \sim \beta$ and $\beta \sim \gamma$ then $\alpha \sim \gamma$;

(3) if $\alpha \sim \beta$ then $\alpha = \beta + \gamma$ where $| \gamma | \ll | \alpha |$ (and $| \gamma | \ll | \beta |$); in other words, the difference between equivalent infinitesimals is an infinitesimal variable of higher order. Conversely, if $\alpha = \beta + \gamma$ where $| \gamma | \ll | \beta |$ then $\alpha \sim \beta$ which means that the addition of a variable of higher order to an infinitesimal results in a variable equivalent to this infinitesimal;

(4) if $\alpha \sim \alpha_1$ and $\beta \sim \beta_1$ then $\lim \frac{x\alpha}{y\beta} = \lim \frac{x\alpha_1}{y\beta_1}$ where x and y are arbitrary variables or numbers; this means that when calculating limits it is allowed to replace infinitesimals occurring in the numerator and in the denominator by equivalent variables.

All these properties can be verified in a similar way. For example, let us justify the fourth property:

$$\frac{x\alpha}{y\beta} = \frac{x\alpha_1}{y\beta_1} \frac{\alpha}{\alpha_1} \frac{\beta_1}{\beta} \quad \text{which implies} \quad \lim \frac{x\alpha}{y\beta} =$$

$$= \lim \frac{x\alpha_1}{y\beta_1} \lim \frac{\alpha}{\alpha_1} \lim \frac{\beta_1}{\beta} = \lim \frac{x\alpha_1}{y\beta_1} \cdot 1 \cdot 1$$

and so the fourth property is true.

9. Important Examples.

1. *The length of an infinitesimal arc is equivalent to that of its chord*, that is $\frac{\overset{\frown}{MN}}{MN} \to 1$ as $N \to M$ (see Fig. 97). The explanation of the property lies in the fact that a small arc is so short that it has no "room" to "crook" noticeably, i.e. to change its direction. Therefore, if we watch these elements "through a microscope" so that

they should be magnified up to finite sizes the chord will be practically indistinguishable from the arc. In more rigorous investigations of this obvious property this fact is sometimes regarded as an axiom
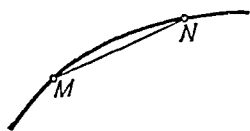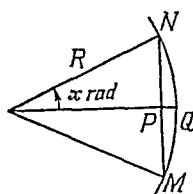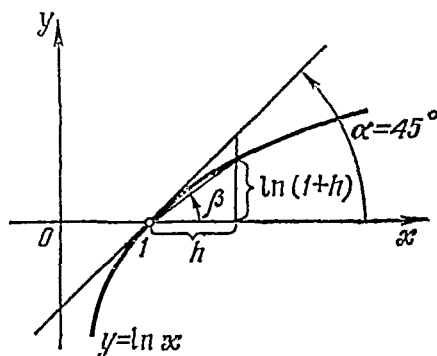
Fig. 97   Fig. 98   Fig. 99

on which the definition of the length of an arc is based. On the other hand, this property is sometimes deduced from other analogous axioms.

2. Applying the above result to an infinitesimal arc of a circle (see Fig. 98) we derive

$$\frac{MN}{\overset{\frown}{MN}} = \frac{2PN}{2\overset{\frown}{QN}} = \frac{2R \sin x}{2Rx} = \frac{\sin x}{x} \xrightarrow[x \to 0]{} 1$$

It was meant that $x > 0$ but the expression $\frac{\sin x}{x}$ does not change when x changes its sign and therefore the sign of x does not matter here. Thus

$$\lim_{x \to 0} \frac{\sin x}{x} = 1 \tag{11}$$

Incidentally, we see that $\sin x < x$ for $x > 0$ (since $MN < \overset{\frown}{MN}$).

3. Now let us consider Fig. 99 which repeats Fig. 41 with some additional lines. We see that $\tan \beta = \frac{\ln(1+h)}{h}$. Now if $h \to 0$ then $\beta \to \alpha = 45°$ and $\tan \beta \to \tan \alpha = 1$, that is

$$\lim_{h \to 0} \frac{\ln(1+h)}{h} = 1 \tag{12}$$

In Fig. 99 it is assumed that $h > 0$ but the same is true for $h < 0$ and $h \to 0$.

Here is an important corollary of formula (12). Since $(1 + h)^{\frac{1}{h}} = e^{\frac{\ln(1+h)}{h}}$, we have, as $h \to 0$,

$$\frac{\ln(1 + h)}{h} \to 1 \quad \text{and} \quad e^{\frac{\ln(1+h)}{h}} \to e^1 = e$$

Thus,

$$\lim_{h \to 0} (1 + h)^{\frac{1}{h}} = e \tag{13}$$

The last limit is sometimes taken as the definition of the number e.

Many other limits can be evaluated by means of the results represented above. Here we also mention that we shall offer a more standard and simple method of evaluating limits in § IV.4.

10. Orders of Smallness. Let α and β be two infinitesimals changing in one and the same process. If β is of the order of α^h then β is said to be an infinitesimal of the kth order relative to α. Here the speed at which α approaches zero serves as some standard with which the speed of β (as β tends to zero) is compared.

Examples. Let $x \to 0$, i.e. let x be an infinitesimal. We shall regard it as a standard. Then if $y = 2x^2$ we see that y is an infinitesimal of the second order (relative to x) since y and x^2 are infinitesimals of the same order; if $z = 4x^3 + x^7$ then z is an infinitesimal of the third order since z and x^3 are of the same order $\left(\frac{z}{x^3} \to 4\right)$.

In general, the sum (or the difference) of infinitesimals of different orders is characterized by the lowest order of the infinitesimals. Namely, the infinitesimal which has the lowest order of smallness is the principal term in such a sum. In other words, all the remaining terms are negligibly small relative to the principal one and the sum is almost completely exhausted by the principal term when the process goes on sufficiently long. Furthermore, if $u = \sqrt{x} - x^2$ then u is of the $\frac{1}{2}$th order and hence u is an infinitesimal of lower order than x, i.e. $\frac{u}{x} \to \infty$ and $|u| \gg |x|$. Generally, if the order of an infinitesimal is lower than unity the infinitesimal is of lower order relative to the standard. Finally, let us take $v = 1 - \cos x$; here v is of the second order since

$$\lim_{x \to 0} \frac{v}{x^k} = \lim_{x \to 0} \frac{1 - \cos x}{x^k} = \lim_{x \to 0} \frac{2 \sin^2 \frac{x}{2}}{x^k} =$$

$$= \lim_{x \to 0} \frac{2 \sin \frac{x}{2} \sin \frac{x}{2}}{x^k} = \lim_{x \to 0} \frac{2 \cdot \frac{x}{2} \cdot \frac{x}{2}}{x^k}$$

(in the passage to the last term we have used the fourth property from Sec. 8) and therefore to receive a finite nonzero limit we should put $k = 2$.

If the standard is changed the order of an infinitesimal may also change and therefore it is absolutely necessary to indicate the standard variable. For instance, the variable $y = x^6$ is an infinitesimal of the sixth order relative to x as $x \to 0$ but it is only of the second order relative to x^3.

11. Comparison of Infinitely Large Variables. The comparison of infinitely large variables is carried out in a way similar to that of comparison of infinitesimals. But there exists a certain difference in the terminology: thus, if $\frac{x}{y} \to 0$ where x and y are infinitely large variables then we say that x is of lower order relative to y and y is of higher order relative to x but the notation $|x| \ll |y|$ or $x = o(y)$ remains (these forms of writing are used in all cases when $\frac{x}{y} \to 0$ even if variables x and y are neither infinitesimal nor infinitely large; we also remark that the notation $x = O(y)$ indicates that the ratio $\frac{x}{y}$ is bounded). All the assertions of Sections 7, 8 and 10 are transferred with some little changes to infinitely large variables.

To conclude the section we point out an obvious property: if $\lim x = 0$, $\lim y = \text{const} \neq 0$ and $\lim |z| = \infty$ then $|x| \ll |y|$ and $|y| \ll |z|$.

§ 4. Continuous and Discontinuous Functions

12. Definition of a Continuous Function. The definition of a continuous function was given in Sec. I.16. Now we are going to discuss it in detail.

Suppose a function $y = f(x)$ is given. Let its argument first take on the value x_0 and then receive an increment Δx, that is let x assume a new value $x = x_0 + \Delta x$ (see Sec. I.22 on this notation). Then the function will also receive an increment

$$\Delta y = y - y_0 = f(x) - f(x_0) = f(x_0 + \Delta x) - f(\dot{x}_0) \qquad (14)$$

The function f is called continuous at the point x_0 (i.e. for the value of the argument equal to x_0) if $\Delta y \to 0$ in a process in which $\Delta x \to 0$ or, in other words, if the increment of the function is an infinitesimal when the increment of the argument is infinitesimal. Otherwise x_0 is called the point of discontinuity of the function f.

Since $f(x) = f(x_0) + \Delta y$ [see formula (14)] the condition $\Delta y \to 0$ is equivalent to the following condition: $f(x) \to f(x_0)$. Further,

writing $x \to x_0$ instead of $\Delta x \to 0$ we arrive at the following equivalent representations of the definition of the continuity:

$$f(x) \xrightarrow[x \to x_0]{} f(x_0) \quad \text{or} \quad \lim_{x \to x_0} f(x) = f(x_0)$$

and, finally,

$$\lim_{x \to x_0} f(x) = f(\lim x) \tag{15}$$

This means that *the limit of a continuous function at a point equals the value which the function assumes when the argument takes on the value of its own limit.**

It should be underlined that the value of a function at a point of its continuity cannot be infinite.

A function which is continuous at each point of an interval is said to be continuous over the interval.

13. Points of Discontinuity. If x_0 is a point of discontinuity of a function f then the value $f(x_0)$ of the function is often undetermined but this fact usually does not play any important role. In these circumstances the limits of the values of the function $f(x)$ as $x \to x_0 - 0$ and $x \to x_0 + 0$ are essentially important (see Sec. 4). These two limits are denoted conditionally as $f(x_0 - 0)$ and $f(x_0 + 0)$, respectively (see Fig. 100).**

It may occur that these limits are finite and $f(x_0 - 0) = f(x_0 + 0)$ though the value $f(x_0)$ may not be defined or it is defined but does not coincide with $f(x_0 \pm 0)$. Such a discontinuity is called removable since if we put $f(x_0) = f(x_0 \pm 0)$ (this value may be called the "true" value of the function $f(x)$ at the point $x = x_0$) then there will no longer exist any discontinuity of $f(x)$. We shall give a simple example: let the function $f(x)$ be defined by the formula

$$f(x) = \frac{\sin x}{x} \tag{16}$$

---

* Recalling the specification pointed out in Sec. 1 we can now formulate the definition of the continuity at the point x_0 as follows: for any given $\varepsilon > 0$ there must exist $\delta > 0$ such that $|x - x_0| < \delta$ should imply $|f(x) - f(x_0)| < \varepsilon$. This definition given by A. L. Cauchy (1789-1857), a prominent French mathematician, is fundamental in books meant for mathematicians.

Defining the continuity of $y = f(x)$ at $x = x_0$ as the condition $\Delta y \to 0$ for $\Delta x \to 0$ we mean that $\Delta y \to 0$ not only for some certain process in which $\Delta x \to 0$ or for several such processes but for all possible processes of this kind. Writing (15) we also mean that it holds for all processes in which $x \to x_0$. Cauchy's definition takes these facts into account automatically since for every process in which $x \to x_0$ there exists a moment from which on $|x - x_0| < \delta$.

** $f(x_0 \mp 0)$ are called, respectively, the limit of $f(x)$ on the left of the point $x = x_0$ (the left-hand limit) and the limit of $f(x)$ on the right of the point $x = x_0$ (the right-hand limit).—*Tr.*

The function is continuous for all $x \neq 0$ but not defined at $x = 0$ since $x = 0$ cannot be substituted into formula (16) because this yields the indeterminate form $\frac{0}{0}$. But if in addition to formula (16) we put $f(0) = 1$ then, by formula (11), the new function $f(x)$ thus obtained will be defined and continuous for all x without exceptions. Thus, we had the removable discontinuity at the point $x = 0$. From
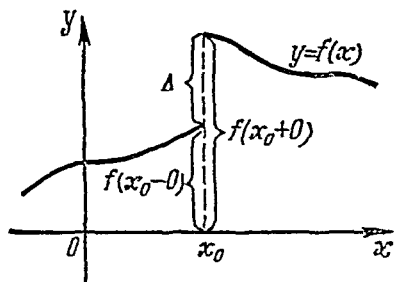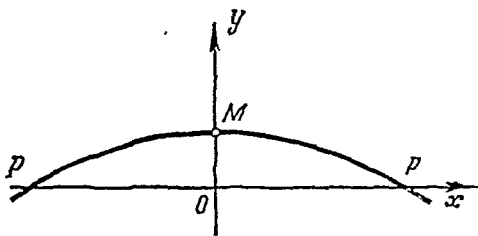
Fig. 100 Fig. 101

the geometrical point of view this means that the curve pp (see Fig. 101) "lacked" one point, i.e. the point M. After the point has been added to the curve it becomes continuous.

If the values $f(x_0 - 0)$ and $f(x_0 + 0)$ are finite but $f(x_0 - 0) \neq f(x_0 + 0)$ then the function f is said to have a **point of discontinuity of the first kind** or, which is the same, has a **finite jump** [the term jump is applied to the value $\Delta = f(x_0 + 0) - f(x_0 - 0)$] or a **jump discontinuity** (see Fig. 100). In case at least one of the values $f(x_0 - 0)$ or $f(x_0 + 0)$ equals infinity we say that the function, in a conditional sense, turns into infinity at the point x_0 (or we say that the graph of the function "travels into infinity"). For instance, the function $f(x) = \frac{1}{x}$ behaves in this way at the point $x_0 = 0$.

In conclusion we remark that in some cases $f(x_0 - 0)$ or $f(x_0 + 0)$ has neither a finite nor an infinite value since a variable may have neither a finite nor an infinite limit. For instance, we have $\frac{1}{x} \to \infty$ as $x \to 0$ and therefore we see that the function $f(x) = \sin \frac{1}{x}$ will infinitely pass from -1 to $+1$ and back to -1 as $x \to 0$ and thus $f(x) = \sin \frac{1}{x}$ has neither a right-hand limit nor a left-hand one (and, in general, it has no limit) at $x = 0$ (see Fig. 102).

Real variable quantities of physical nature have discontinuities when a certain kind of action is suddenly applied or switched off,

when there is a transition from one medium into another (through the interface between the two media), when the law of functional relationship between the quantities suddenly changes etc.

See, for example, Fig. 103. There is a graph of variations of the electric current flow i plotted against time t there which corresponds
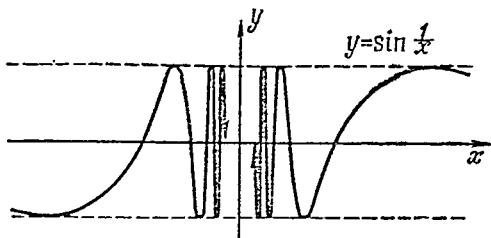
Fig. 102

to the process of transmitting the letter "a" in the Morse code by radio (the signal "dot—dash"). Thus Fig. 103 shows the dependence of i on t. As is seen, we have here a function with four points of discontinuity and at each of these points there is a finite jump corresponding to switching on or switching off the constant emf (electromotive force) in the circuit. It is important to pay attention to the
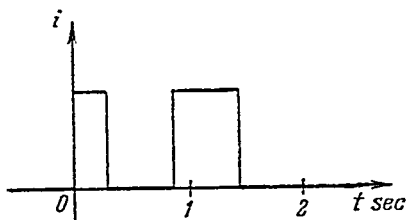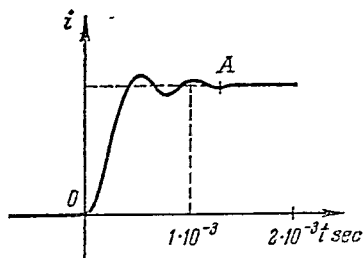
Fig. 103   Fig. 104

fact that if we analysed this phenomenon more carefully (and for this purpose took a much larger time scale for the t-axis) then we should see that in reality the growth of the current is similar to the one shown in Fig. 104. Since there is always a certain inductance in the circuit the current increases continuously (although very fast) and therefore in real circumstances there must be no discontinuity of the function i (t) at all! In some cases when, for instance, a pulse lasts a very short time it may be important to take into account the continuity of this transient process (for example, corresponding to the part OA of the graph). But when the transient process is of no importance

it is simpler to schematize the process and consider the function i (t) discontinuous according to Fig. 103 if this does not lead to noticeable mistakes. Thus, one and the same function i (t) of a real "physical" nature may be regarded as continuous or discontinuous depending on whether we intend to take into account the transient process or not. In the case of passage from one medium into another an analogous role is played by the processes near the interface between the media which may or may not be taken into account.

If an elementary function is considered then, as it will be shown in Sec. 14, it can have a discontinuity at $x = x_0$ if and only if the substitution of $x = x_0$ into the function yields an expression of types $\frac{a}{0}$, ln 0 or 0^0 in the expression of f (x_0) or in some part of this expression. G. H. Hardy (1877-1947), an English mathematician, showed that in case such a situation takes place the limits f $(x_0 - 0)$ and f $(x_0 + 0)$, finite or infinite, should exist provided the function f is defined, respectively, on the left or on the right of x_0. Only in case the expressions sin ∞ and cos ∞ enter into the representation of f (x_0) there may be an exception to this rule.

If a function f is defined only on one side of x_0, for example, on the right, then it may happen that only f $(x_0 + 0)$ ("the end-point value") exists while f $(x_0 - 0)$ does not. The limits f $(-\infty)$ and f $(+\infty)$ may also be regarded as end-point values.

14. Properties of Continuous Functions.

1. *The sum of two continuous functions is a continuous function.* Virtually, if the functions f_1 (x) and f_2 (x) are continuous and f $(x) = f_1$ $(x) + f_2$ (x) then, as $x \to x_0$,

$$\lim f (x) = \lim [f_1 (x) + f_2 (x)] = \lim f_1 (x) + \lim f_2 (x) =$$
$$= f_1 (x_0) + f_2 (x_0) = f (x_0)$$

which implies the continuity of the function f (x) (see Sec. 12). We point out that in the above proof we first used a property of limits (see Sec. 5) and then the continuity of the functions f_1 and f_2. Similar application of other properties of limits implies the following:

a sum or a difference or a product of an arbitrary number of continuous functions is also a continuous function;

a ratio of two continuous functions is a function which is continuous everywhere except the points where the denominator equals zero. The ratio either approaches infinity or becomes an indeterminate form of type $\frac{0}{0}$ at those points where the denominator vanishes. Therefore the continuity does not hold in either case.

2. *A composite function formed by means of continuous functions is a continuous function.* Indeed, if the functions z (y) and y (x) are continuous and x assumes an infinitesimal increment then, by the continuity of the second function, the increment of y will be infi-

nitesimal too and therefore the increment of z will be also infinitesimal by the continuity of the first function; thus, the composite function $z(x)$ will be continuous.

The first two properties enable us to make the following conclusion concerning the continuity of elementary functions. Reviewing the basic elementary functions (see Sec. I.18 and § I.4) we see that among them only $y = x^{-m}$ has a discontinuity for $-m < 0$ at $x = 0$ (in this case the form $\frac{1}{0}$ appears), $y = \log_a x$ has a discontinuity at $x = 0$ (this yields log 0) and $y = \tan x$ at $x = \pm \frac{\pi}{2}, \pm \frac{3\pi}{2}, \ldots$ (this results in $\frac{\sin \frac{\pi}{2}}{\cos \frac{\pi}{2}} = \frac{1}{0}$). When composite functions and algebraic combinations are composed of basic functions then, according to properties 1 and 2 and Sec. 15 (1), the new points of discontinuity may appear if and only if the expressions of the form $\frac{a}{0}$ and 0^0 occur. This proves the assertion stated in the end of Sec. 13.

Hence, if $f(x)$ is an elementary function then $\lim\limits_{x \to x_0} f(x)$ simply equals $f(x_0)$ provided there are no "dangerous" expressions of the form $\frac{a}{0}$, ln 0 and 0^0 in the expression of $f(x)$; for example,

$$\lim_{x \to 1} \frac{\ln(1 + \sin x)}{x^2 - 2x} = \frac{\ln(1 + \sin 1)}{1^2 - 2 \cdot 1} = -\ln(1 + \sin 1) =$$
$$= -\ln 1.8415 = -0.6105$$

(the values of sin 1 and of ln 1.8415 are taken from tables). This rule of determining a limit remains true for the operations on infinities provided after substituting the limit of the argument the indeterminate forms $\frac{0}{0}$, $\infty - \infty$, $\frac{\infty}{\infty}$, $0 \cdot \infty$ (which were mentioned in Sec. 5), 0^0, 1^∞, ∞^0 [which will be discussed in Sec. 15 (1)] and expressions of the form sin ∞ do not appear. For example,

$$\lim_{x \to +0} \left(\frac{\ln x}{x} + \cos x \right) = \frac{\ln(+0)}{+0} + \cos(+0) = \frac{-\infty}{+0} + 1 = -\infty,$$
$$\lim_{x \to +0} (x^{\frac{1}{x}} + x^{-\frac{1}{x}}) = (+0)^{+\infty} + (+0)^{-\infty} = 0 + \infty = \infty$$

More on the indeterminate forms see in § 3 and § IV.4.

3. As it was indicated in Sec. I.16, the graph of a continuous function $y = f(x)$ defined in an interval $a \leqslant x \leqslant b$ consists of one component. The consideration of such a graph (see, for example, Fig. 105)

shows that a *continuous function defined over a finite interval* (*inclu-ding its end-points*) *is bounded and attains its least* (in the algebraic sense) *value* (at $x = a$ in Fig. 105) *and its greatest* (at $x = c$) *value.* These values are denoted, respectively, as $\min\limits_{a \leqslant x \leqslant b} f(x)$ and $\max\limits_{a \leqslant x \leqslant b} f(x)$ (abbreviations of "maximum" and "minimum"). Besides, *such a func-tion takes on all the intermediate values between $f(a)$ and $f(b)$, each value being taken at least one time.* For example, the value $y = q$ is assumed only one time in Fig. 105, i.e. at $x = \delta$, whereas the value $y = p$ is taken three times at the points $x = \alpha$, β and γ. Further,
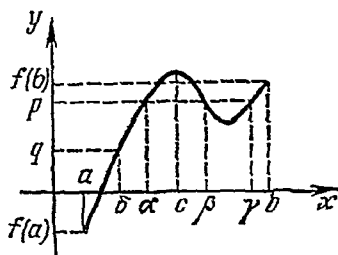
Fig. 105

Fig. 106

if $f(x)$ is positive for some value $x = x_0$ it remains positive at all points x ying close enough to x_0. Finally, turning back to Figs. 25 and 26 we see that *if a continuous function is monotonic then its inverse func-tion is also continuous* (*and monotonic*).

Here we shall restrict ourselves to the visual considerations con-cerning these properties. Their rigorous proof does not appear simple in the general case; this proof can be found, for instance, in [14].

It should be pointed out that a continuous function is not necessa-rily smooth, that is having a graph with a certain tangent at its every point (a smooth function is shown in Fig. 105). On the contrary, as it is shown in Fig. 106, it may happen to be **piecewise smooth**, that is to have a broken graph consisting of several smooth arcs. Even some more complicated cases are possible but we are not going to treat them in our course. Some more detailed comments on this question will be given in Sec. IV.3.

15. Some Applications.

1. *Limits of Composite Exponential Expressions* ("Power-Exponen-tial" Expressions). Let us consider the expression x^y where $x \to a$, $y \to b$ and $x > 0$. Let us represent x^y in the form $x^y = (e^{\ln x})^y = e^{y \ln x}$. But $\ln x \to \ln a$ by the continuity of the logarithmic function and hence $y \ln x \to b \ln a$ (as a limit of a product). There-fore $\exp(y \ln x) \to \exp(b \ln a)$ (by the continuity of the exponen-tial function). Thus, $x^y \to e^{b \ln a} = (e^{\ln a})^b = a^b$ or, in other words,

$$\lim x^y = (\lim x)^{\lim y}$$

and so we see that it is permissible to pass to the limit in the expression x^y. The exception to this rule is the case when the product $b \ln a$ becomes an indeterminate form, i.e. it has the form $0 \cdot \infty$. This may occur if

$$\ln a = 0; \quad b = \infty, \quad \text{i.e.} \quad a = 1 \quad \text{and} \quad a^b = 1^\infty;$$

$$-\ln a = \infty; \quad b = 0, \quad \text{i.e.} \quad a = \infty \quad \text{and} \quad a^b = \infty^0;$$

$$\ln a = -\infty; \quad b = 0, \quad \text{i.e.} \quad a = 0 \quad \text{and} \quad a^b = 0^0$$

Hence we just have the three types of the indeterminate forms mentioned in Sec. 14.

One may sometimes think that there must be $1^\infty = 1$ since "unity to any power equals unity". But 1^∞ is not at all unity to a certain

Fig. 107

finite power but only the abbreviated notation for a limit of an expression of the form x^y where $x \to 1$ and $y \to \infty$. Suppose, for example, that $x \to 1 + 0$, that is $x > 1$. Then the expression x^y "has a certain tendency to approach 1" (since $1^y = 1$) and at the same time it "wants to tend to infinity" (since $x > 1$ and $x^\infty = \infty$ because if we raise a constant number larger than unity to an infinitely increasing power we shall arrive at an infinitely large variable). Therefore, these two tendencies "act" upon the expression in the opposite directions and hence the result may be different in different problems depending on which of these tendencies "wins". For example, in case (13) the limit turned out to be equal to e whereas the immediate substitution of $h = 0$ yields 1^∞. In this case we see that the tendencies are equal in a certain sense; they are in a state of "balance". Similar conclusions may be derived in connection with other indeterminate forms.
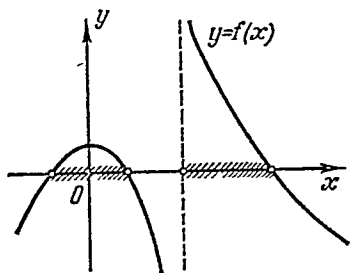
2. *Solving Inequalities*. Let a function $f(x)$ be considered on an interval (a, b) (in particular, the interval may be the whole x-axis) and let it be necessary to solve the inequality

$$f(x) > 0 \qquad (17)$$

that is to determine all the values of x for which it holds. In the geometrical sense this means that we must find such regions on the x-axis where the graph of the function $y = f(x)$ lies over the x-axis (such regions are shaded in Fig. 107). But it must be stressed that in this problem we regard the function $f(x)$ as known whereas its graph may be unknown.

To solve inequality (17) let us mark all the *zeros of the function f* (that is the points where f vanishes) on the interval (a, b) and also all the points of discontinuity (there are three zeros and one point of discontinuity in Fig. 107). The interval is divided into several parts by these points (five parts in Fig. 107). Since we have taken into account all the points of discontinuity of the function it is continuous inside each of these parts. Besides the function does not vanish inside the parts because we have reckoned all its zeros. Thus, the function $y = f(x)$ retains its sign inside each of the parts (see
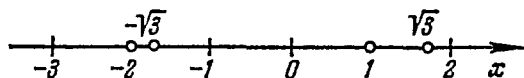
Fig. 108

property 3 in Sec. 14). Now in order to determine this sign it is sufficient to determine the sign of the function at any point inside the considered part of the interval. After this operation we choose those parts in which the function is positive and thus the inequality (17) will be solved.

Example. Let us solve the inequality $\dfrac{x^3 + 3x^2 - 4}{x^2 - 3} > 0$.

The numerator equals zero when $x = 1$ and therefore it is divisible by $x - 1$. This implies that the expression

$$y = \frac{x^3 + 3x^2 - 4}{x^2 - 3} = \frac{(x-1)(x^2 + 4x + 4)}{x^2 - 3} = \frac{(x-1)(x+2)^2}{x^2 - 3}$$

must be positive for the values of x which we are interested in. Hence, the function is defined over the whole x-axis except $x = \pm\sqrt{3}$ and has two zeros ($x = 1$ and $x = -2$) and two points of discontinuity ($x = \pm\sqrt{3}$). These points break the x-axis into five parts (see Fig. 108). Now we choose a point in each of the intervals, substitute these values into the last fraction and determine the signs of the fraction inside the intervals (the numerical values themselves do not matter and only their signs are essential). Thus we receive the table

x	-3	-1.9	0	1.1	2
y	$-$	$-$	$+$	$-$	$+$

Thus, the solution of the inequality is a totality consisting of two intervals:

$$-\sqrt{3} < x < 1 \quad \text{and} \quad \sqrt{3} < x < \infty$$

Derivatives, Differentials, Investigation of the Behaviour of Functions

§ 1. Derivative

1. Some Problems Leading to the Concept of a Derivative. We come to the notion of a derivative, one of the most important notions in mathematics, when investigating the rate of change of a function.

For example, let us turn to the notion of the velocity (rate) at a given instant of a rectilinear motion of a material point. A material point is understood in physics as a material body such that it is permissible to neglect its geometrical sizes while investigating the state of the body under some concrete conditions. In different circumstances a particle of a substance, or an airplane, or a heavenly
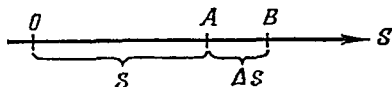
Fig. 109

body etc. may sometimes be regarded as a point. Let a material point move along the s-axis from left to right. In the general case the motion may be non-uniform, that is the velocity of the motion may be variable. The law of motion is expressed mathematically as a dependence of the coordinate s on time t: $s = f(t)$. Since the velocity is variable the ratio of the distance passed over to the time taken represents the **average velocity** only. As for the "true" velocity, that is the velocity at a given instant, it can be obtained by means of the following procedure. Let the moving point occupy the position A (see Fig. 109) at an instant t. Suppose during the period of time Δt (see Sec. I.22 on this notation) the moving point transits to the position B, the distance Δs being passed over. Then

$$s = f(t), \quad s + \Delta s = f(t + \Delta t)$$

i.e. $\Delta s = f(t + \Delta t) - f(t)$. Hence, the ratio $v_{av} = \dfrac{\Delta s}{\Delta t}$ (which is the distance passed over per unit of time taken) is the average

velocity of motion during the time period Δt from t to $t + \Delta t$. Now *the instantaneous velocity of motion at time t is obtained as the limit of the average velocity in the process of decreasing the interval Δt unlimitedly*, that is

$$v_{inst} = \lim_{\Delta t \to 0} v_{av} = \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t} = \lim_{\Delta t \to 0} \frac{f(t + \Delta t) - f(t)}{\Delta t} \tag{1}$$

It is also said that the instantaneous velocity (that is the velocity at a given instant, the true velocity) is the average velocity during an infinitesimal interval of time ("element" of time) or that the instantaneous velocity is the ratio of an infinitesimal distance to an infinitesimal time interval. Both definitions briefly express the meaning of the general definition (1).

The rate of a physical process is not in all cases represented by the distance passed over related to the unit of the time taken. Let us con-

Fig. 110

sider, for example, the process of filling a vessel. In this case the dependence $V = f(t)$ of the volume already filled on time t expresses the law of the process of filling. The average rate of filling during the interval of time from t to $t + \Delta t$ is represented by the ratio

$$w_{av} = \frac{\Delta V}{\Delta t} = \frac{f(t + \Delta t) - f(t)}{\Delta t}$$

whereas the limit

$$w_{inst} = \lim_{\Delta t \to 0} w_{av} = \lim_{\Delta t \to 0} \frac{\Delta V}{\Delta t} = \lim_{\Delta t \to 0} \frac{f(t + \Delta t) - f(t)}{\Delta t} \tag{2}$$

serves as the instantaneous rate, i.e. the rate of filling at time t. Thus we have arrived at an expression similar to (1).

But we can understand the velocity, the rate, even in a wider sense relating the change of a quantity not to the unit of time but to the unit of some other quantity. For example, let us consider the notion of the linear density of a material line, that is of a body such that, under given concrete conditions, it is permissible to take into account only its size in one-dimensional extent (the longitudinal size) neglecting the cross-section sizes. At the same time we do not neglect its mass. If this line ("thread") is homogeneous its linear density is equal to the ratio of its mass to its length. In case the thread is non-homogeneous its linear density is different at different points. Let us reckon the distance from one of the ends of the thread (see Fig. 110) and let the mass of the part of the thread corresponding to the distance s be equal to $M = f(s)$. If now some additional
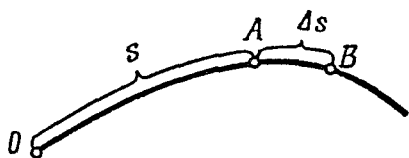
distance Δs is passed the ratio

$$\rho_{av} = \frac{\Delta M}{\Delta s} = \frac{f(s + \Delta s) - f(s)}{\Delta s}$$

represents the average linear density of the thread corresponding to the part AB. The limit

$$\rho = \lim_{\Delta s \to 0} \rho_{av} = \lim_{\Delta s \to 0} \frac{\Delta M}{\Delta s} = \lim_{\Delta s \to 0} \frac{f(s + \Delta s) - f(s)}{\Delta s} \qquad (3)$$

now gives the linear density of the thread at a point (namely, at the point A). We may say that ρ is the rate (velocity) of change, of the mass of the thread, i.e. the change of the mass per unit of distance passed.

2. **Definition of Derivative.** From the mathematical point of view expressions (1), (2) and (3) are quite similar. This enables us to state the following definition. Let a function $y = f(x)$ be given. Then the rate of its change related to the unit of change of the argument x is equal to

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

This rate (velocity) is called **the derivative of the variable (function) y with respect to the variable (argument) x**; in other words, *the derivative is the limit of the ratio of the increment of the function to the increment of the argument taken in the process when the increment of the argument approaches zero.* Since this rate has, in general, different values for different values of x the derivative itself is a new function of x. This new function is designated as $y' = f'(x)$.

Hence, in the examples of Sec. 1 the velocity of motion is equal to the derivative of the distance passed with respect to the time, i.e. $v = s_t$ (the subscript t in the expression s_t indicates that the derivative is taken with respect to the variable t) etc.

For example, let us compute the derivative of the function $y = ax^2$. Increasing the argument by an increment Δx we receive the new value of the argument $x + \Delta x$ and the new value of the function $y + \Delta y = a(x + \Delta x)^2$ since $x + \Delta x$ should be substituted for x into the expression of the function. Thus,

$$\Delta y = a(x + \Delta x)^2 - ax^2 = 2ax\,\Delta x + a(\Delta x)^2$$

This implies

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{2ax\,\Delta x + a(\Delta x)^2}{\Delta x} = \lim_{\Delta x \to 0} (2ax + a\,\Delta x) = 2ax$$

Note that in the latter passage to the limit only Δx varied as $\Delta x \to 0$ whereas x was considered to be constant. The result thus obtained can be written in the form $(ax^2)' = 2ax$.

We leave it to the reader to verify that $(ax^3)' = 3ax^2$, $(ax)' = a$ and the like. We particularly note here that $x' = 1$.

3. **Geometrical Meaning of Derivative.** Let us consider the graph of a function $f(x)$ (see Fig. 111). We see that $\frac{\Delta y}{\Delta x} = \frac{PN}{MP} = \tan \beta$, i.e. the ratio is equal to the slope of the secant mm. If $\Delta x \to 0$ then the secant turns round the point M and tends to the position of the
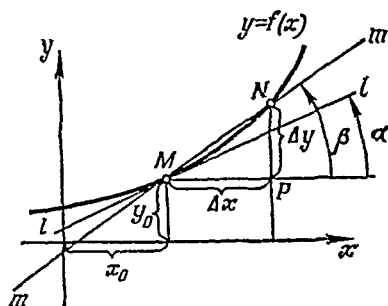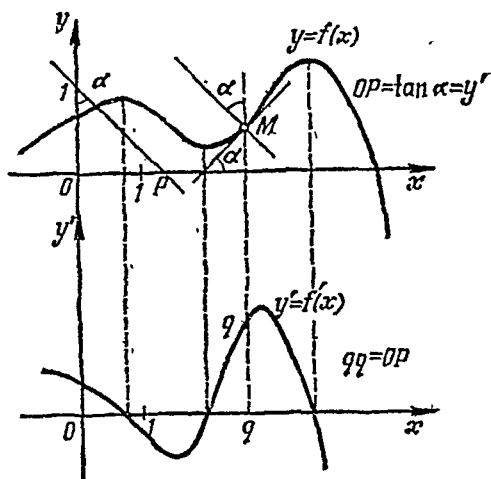
Fig. 111   Fig. 112

tangent ll in the limit process since *the tangent occupies the limiting position of the secant when the points of intersection merge*. (This obvious property which we have already used is in fact nothing but the definition of a tangent.) Therefore

$$y_0' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim \tan \beta = \tan \alpha \quad [y_0' = f'(x_0)] \qquad (4)$$

that is *the geometrical meaning of the derivative of a function is that it is equal to the slope of the tangent*. By formula (II.21) it is easy now to put down the equation of the tangent ll:

$$y - y_0 = y_0'(x - x_0) \qquad (5)$$

where x_0 and y_0 are the coordinates of the point of tangency, x and y are the moving coordinates of the point on the tangent straight line.

Similarly, the equation of the **normal** to the curve, that is of the line perpendicular to the tangent at the point of tangency, has the form $y - y_0 = -\frac{1}{y_0'}(x - x_0)$ (see problem 5 in Sec. II.9).

In Sec. I.26 we said that the angle between two curves at the point of their intersection was defined as the angle between the tangents to the curves at that point, and therefore we are able now to determine the angle by means of formula (II.23) since we know how to

determine the tangents. Note that the angle may turn out to be zero in case these curves are tangent to each other, i.e. when their tangents coincide.

When the graph of a function $y = f(x)$ is given the geometrical meaning of the derivative makes it possible to indicate the slope of the tangent to the graph and this enables us to draw immediately a sketch of the graph of the derivative (see Fig. 112). For more accurate "graphical computation of the derivative" it is necessary to draw tangents to the given graph and measure their slopes. It turns out that it is practically simpler to draw normals to the graph by means of a shiny (metallic) ruler and to measure their slopes with
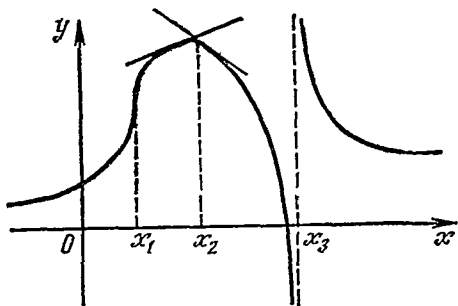
Fig. 113

respect to the y-axis which is just the same. One of these procedures is shown in Fig. 112. We apply the ruler perpendicularly to the plane of the graph to one of its points, e.g. to the point M, and turn the ruler in such a way that the reflection of the graph in the ruler should prolong the graph without a break at M. In this position the ruler will lie exactly along the normal to the graph at M. Then we draw a straight line passing through the point $(0; 1)$ and parallel to the normal thus constructed. In this way we get the line segment OP which is then transferred to the position qq. After a number of such procedures are carried out we obtain a rather accurate graph of the derivative.

While discussing the geometrical meaning of the derivative [formula (4)] we supposed that both variables x and y were dimensionless and that the scale was the same for both axes. But this is not always the case in practical problems. It follows from formula (I.10) that in the general case we must write $y'_0 = \frac{l_x}{l_y} \tan \alpha$. Thus in the general case the derivative is also equal to the slope of the tangent.

Note that if the derivative y' approaches infinity for some value of x (for $x = x_1$ in Fig. 113) then the tangent at the corresponding

point of the graph has the slope equal to infinity, that is the tangent line is parallel to the y-axis. If the derivative has a jump discontinuity at a point then the tangent turns jumpwise, i.e. the graph is broken at the point (see the point $x = x_2$ in Fig. 113). In case the function approaches infinity the derivative may also turn into infinity (see the point $x = x_3$ in Fig. 113).

4. Basic Properties of Derivatives.

1. *The derivative of a constant equals zero.* (The property is obviously interpreted as the fact that the velocity of a body in a state of rest is equal to zero.) The formal proof of the property looks as follows: if $y = C = \text{const}$ then

$$\Delta y = C - C = 0, \quad \frac{\Delta y}{\Delta x} = 0,$$

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} 0 = 0$$

2. *The derivative of a sum is equal to the sum of the derivatives of the summands.* Indeed, if $y(x) = u(x) + v(x)$ then $y(x + \Delta x) = u(x + \Delta x) + v(x + \Delta x)$ and

$$\Delta y = y(x + \Delta x) - y(x) = [u(x + \Delta x) + v(x + \Delta x)] -$$
$$- [u(x) + v(x)] = [u(x) + \Delta u + v(x) + \Delta v] -$$
$$- [u(x) + v(x)] = \Delta u + \Delta v$$

that is *the increment of a sum is equal to the sum of the increments:* $\Delta(u + v) = \Delta u + \Delta v$. Hence, it follows that

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta u + \Delta v}{\Delta x} = \lim_{\Delta x \to 0} \left(\frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta x} \right) =$$
$$= \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} + \lim_{\Delta x \to 0} \frac{\Delta v}{\Delta x} = u' + v'$$

(in the deduction we have used the fact that the limit of a sum is equal to the sum of the limits) which is the required proof. This property can be rewritten in a different way as

$$(u + v)' = u' + v'$$

We have taken a sum of two summands. It is clear that the same is true for an arbitrary number of summands. Similarly, the increment of a difference is equal to the difference of the increments and the derivative of a difference is equal to the difference of the derivatives.

Example. $(x^3 - 3x^2 + x + 5)' = (x^3)' - (3x^2)' + (x)' + (5)' = 3x^2 - 6x + 1 + 0 = 3x^2 - 6x + 1$ (see the end of Sec. 2).

3. *A constant factor can be taken outside the derivative sign*, i.e.

$(Cu)' = Cu'$ (where $C = $ const). Virtually, if $y = Cu$ then
$$\Delta y = y\,(x + \Delta x) - y\,(x) = Cu\,(x + \Delta x) - Cu\,(x) =$$
$$= C\,[u\,(x + \Delta x) - u\,(x)] = C\,\Delta u$$

in other words. *if a function is multiplied by a constant its increment is multiplied by the same constant*: $\Delta\,(Cu) = C\,\Delta u$.
Hence.

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{C\,\Delta u}{\Delta x} = C \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} = Cu'$$

4. *Formula for the Derivative of a Product.* Let $y = uv$. Then
$\Delta y = (u + \Delta u)\,(v + \Delta v) - uv = (\Delta u)\,v + u\,\Delta v + \Delta u\,\Delta v$ which implies

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{(\Delta u)\,v}{\Delta x} + \lim_{\Delta x \to 0} \frac{u\,\Delta v}{\Delta x} + \lim_{\Delta x \to 0} \frac{\Delta u\,\Delta v}{\Delta x} =$$

$$= \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} v + \lim_{\Delta x \to 0} u\,\frac{\Delta v}{\Delta x} + \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x}\,\frac{\Delta v}{\Delta x}\,\Delta x = u'v + uv' + u'v'\cdot 0$$

Thus,

$$(uv)' = u'v + uv' \tag{6}$$

For example,
$$[(3x^2 + 5x)\,(4x^2 - 6)]' = (3x^2 + 5x)'\,(4x^2 - 6) +$$
$$+ (3x^2 + 5x)\,(4x^2 - 6)' = (6x + 5)\,(4x^2 - 6) +$$
$$+ (3x^2 + 5x)\,8x = 48x^3 + 60x^2 - 36x - 30$$

From formula (6) we can easily deduce the formula for the derivative of a product of several factors. For example,
$$(uvw)' = [(uv)\,w]' = (uv)'\,w + (uv)\,w' =$$
$$= (u'v + uv')\,w + uvw' = u'vw + uv'w + uvw'$$

The formula for the derivative of a product of an arbitrary number of factors looks quite similar. We note that property 3 can be easily deduced from formula (6) by putting $v = C$.

5. *Formula for the Derivative of a Quotient.* Let $y = \dfrac{u}{v}$. Then

$$\Delta y = \frac{u + \Delta u}{v + \Delta v} - \frac{u}{v} = \frac{(\Delta u)\,v - u\,\Delta v}{v\,(v + \Delta v)}$$

From this, representing Δv which enters into the denominator in the form $\dfrac{\Delta v}{\Delta x}\,\Delta x$, we obtain

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\dfrac{\Delta u}{\Delta x}\,v - u\,\dfrac{\Delta v}{\Delta x}}{v\left(v + \dfrac{\Delta v}{\Delta x}\,\Delta x\right)} = \frac{u'v - uv'}{v\,(v + v'\cdot 0)}$$

Thus,

$$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2} \tag{7}$$

For example,

$$\left(\frac{5x^2}{3x^2+4}\right)' = \frac{(5x^2)'\,(3x^2+4) - (5x^2)\,(3x^2+4)'}{(3x^2+4)^2} =$$

$$= \frac{10x\,(3x^2+4) - 5x^2 \cdot 6x}{(3x^2+4)^2} = \frac{40x}{(3x^2+4)^2}$$

6. *The Derivative of a Composite Function.* Let $y = f(u)$, $u = \varphi(x)$ and let y be regarded as a composite function of x. If x receives an increment Δx then the intermediate variable u receives an increment Δu and therefore y receives an increment Δy too. We have

$$\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x} \quad {}^* \tag{8}$$

Now let $\Delta x \to 0$. Then $\frac{\Delta u}{\Delta x} \to u'_x$ and hence $\Delta u = \frac{\Delta u}{\Delta x}\,\Delta x \to u'_x \cdot 0 =$

$= 0$. Therefore $\frac{\Delta y}{\Delta u} \to y'_u$. Passing to the limit in formula (8) we obtain

$$y'_x = y'_u u'_x \tag{9}$$

The last formula may be rewritten as

$$[f(\varphi(x))]' = f'(\varphi(x))\,\varphi'(x) \tag{10}$$

In the case when a composite function is formed by means of a greater number of intermediate stages the derivative is computed in the same way. Thus, if $y = y(u)$, $u = u(v)$ and $v = v(x)$ then $y'_x = y'_u \cdot u'_v \cdot v'_x$.

For instance, let $y = (x^2 - 5x + 3)^3$. Then we can denote $y = u^3$ where $u = x^2 - 5x + 3$, and by formula (9) we get

$$y'_x = y'_u \cdot u'_x = (u^3)'_u \cdot (x^2 - 5x + 3)'_x = 3u^2\,(2x - 5) =$$
$$= 3\,(x^2 - 5x + 3)^2\,(2x - 5)$$

which is, of course, simpler than removing the brackets! In practical computations there is no need to write down all this in such a detailed manner. For instance, the former calculations can be put down as follows:

$$[(x^2 - 5x + 3)^3]' = 3\,(x^2 - 5x + 3)^2\,(x^2 - 5x + 3)' =$$
$$= 3\,(x^2 - 5x + 3)^2\,(2x - 5)$$

---

* Formula (8) is valid, of course, only if $\Delta u \neq 0$. But in case $\Delta u = 0$ it is also easy to prove formula (10).—*Tr.*

We use here formula (10), of course. After some practice one can write the result immediately without intermediate transformations. For this purpose it is advisable to remember the formulas for the derivatives in the form $(u^2)' = 2uu'$, $(u^3)' = 3u^2u'$ and the like (the derivatives are taken with respect to x).

7. *The Derivative of an Inverse Function.* Suppose the equality $y = y(x)$ defines the inverse relation $x = x(y)$ (see Sec. I.21) for which we can determine the derivative x'_y. Then it is easy to compute the derivative of the original function $y(x)$. Indeed, $\frac{\Delta y}{\Delta x} = \frac{1}{\frac{\Delta x}{\Delta y}}$, which implies, as $\Delta x \to 0$ and $\Delta y \to 0$,

$$y'_x = \frac{1}{x'_y} \tag{11}$$

For example, let $y = \sqrt[3]{x}$ which yields $x = y^3$. Then

$$y'_x = \frac{1}{x'_y} = \frac{1}{(y^3)'_y} = \frac{1}{3y^2} = \frac{1}{3\sqrt[3]{x^2}}$$

8. *The Derivative of an Implicit Function.* If a function is determined in an implicit form $F(x, y) = 0$ (see Sec. I.20) then to compute the derivative y'_x one should simply equate the derivatives of the left-hand side and of the right-hand side of the latter relation taking into account that y is a function of x which turns the relation into an identity. Generally, it is permissible to equate the derivatives of both sides of an equality if and only if the equality is an identity (but not an equation!).

For instance, let us take

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{12}$$

Then

$$\left(\frac{x^2}{a^2}\right)'_x + \left(\frac{y^2}{b^2}\right)'_x = (1)'_x, \text{ i.e. } \frac{2x}{a^2} + \frac{2yy'}{b^2} = 0 \tag{13}$$

Computing the derivative of the second summand we have used property 6: $\left(\frac{y^2}{b^2}\right)'_x = \frac{1}{b^2}(y^2)'_y \cdot y'_x = \frac{1}{b^2} 2yy'$. Thus, (13) implies

$$y' = -\frac{b^2 x}{a^2 y} \tag{14}$$

5. Derivatives of Basic Elementary Functions.

1. *The Derivative of Sine.* Let $y = \sin x$. If the argument changes and becomes equal to $x + \Delta x$ then the function becomes equal to $\sin(x + \Delta x)$. This implies

$$\Delta y = \sin(x + \Delta x) - \sin x = 2\sin\frac{\Delta x}{2} \cdot \cos\left(x + \frac{\Delta x}{2}\right);$$

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\left[2 \sin \frac{\Delta x}{2} \cdot \cos \left(x + \frac{\Delta x}{2} \right) \right]}{\Delta x} =$$

$$= \lim_{\Delta x \to 0} \frac{\sin \frac{\Delta x}{2}}{\frac{\Delta x}{2}} \lim_{\Delta x \to 0} \cos \left(x + \frac{\Delta x}{2} \right) = 1 \cdot \cos x$$

[see formula (III.11)]. Hence,

$$(\sin x)' = \cos x \tag{15}$$

2. We leave it to the reader to verify in an analogous way that

$$(\cos x)' = -\sin x \tag{16}$$

3. *The derivative of tangent* is calculated by formula (7):

$$(\tan x)' = \left(\frac{\sin x}{\cos x} \right)' = \frac{(\sin x)' \cos x - \sin x (\cos x)'}{\cos^2 x} =$$

$$= \frac{\cos x \cdot \cos x - \sin x (-\sin x)}{\cos^2 x} = \frac{1}{\cos^2 x}$$

4. Similarly, we can verify that $(\cot x)' = -\dfrac{1}{\sin^2 x}$.

5. *The Derivative of Arc Sine.* Take $y = \arcsin x$. Then $x = \sin y$ and, by formula (11),

$$y'_x = \frac{1}{x'_y} = \frac{1}{(\sin y)'_y} = \frac{1}{\cos y} = \frac{1}{\pm \sqrt{1 - \sin^2 y}} = \frac{1}{\sqrt{1 - x^2}}$$

We have written $+$ in front of the radical sign because the values of $\arcsin x$, as is well known, are taken in the interval $-\dfrac{\pi}{2} \leqslant \arcsin x \leqslant \dfrac{\pi}{2}$ which corresponds to non-negative values of $\cos y \geqslant 0$. Thus,

$$(\arcsin x)' = \frac{1}{\sqrt{1 - x^2}}$$

6. We verify similarly that

$$(\arccos x)' = -\frac{1}{\sqrt{1 - x^2}} \tag{17}$$

using the inequality $0 \leqslant \arccos x \leqslant \pi$. The resemblance between the last two results is explained by the formula

$$\arcsin x + \arccos x \equiv \frac{\pi}{2} \tag{18}$$

which can be deduced in the following way: if we denote $\sin \alpha = x$ then

$$\cos \left(\frac{\pi}{2} - \alpha \right) = x$$

and these formulas yield $\alpha = \text{arc sin } x$ and $\frac{\pi}{2} - \alpha = \text{arc cos } x$. The addition of the last results implies (18).

7. *The Derivative of Arc Tangent.* Let $y = \text{arc tan } x$. Then $x = \tan y$. Using the formulas for the derivatives of an inverse function and of the tangent we obtain

$$y'_x = \frac{1}{x'_y} = \frac{1}{\dfrac{1}{\cos^2 y}} = \cos^2 y = \frac{1}{1 + \tan^2 y} = \frac{1}{1 + x^2}$$

Thus,

$$(\text{arc tan } x)' = \frac{1}{1 + x^2}$$

8. *The Derivative of a Logarithmic Function.* Take $y = \ln x$. Then putting $h = \frac{\Delta x}{x}$ in formula (III.12) we see that

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\ln(x + \Delta x) - \ln x}{\Delta x} = \lim_{\Delta x \to 0} \frac{\ln\left(1 + \frac{\Delta x}{x}\right)}{\frac{\Delta x}{x} x} = \frac{1}{x}$$

Therefore

$$(\ln x)' = \frac{1}{x}$$

Applying formula (I.14) and taking into account that $\ln a = \text{const}$ we receive

$$(\log_a x)' = \left(\frac{\ln x}{\ln a}\right)' = \frac{1}{\ln a}(\ln x)' = \frac{1}{x \ln a}$$

9. *The Derivative of an Exponential Function.* If $y = a^x$ then

$$x = \log_a y \quad \text{and} \quad y'_x = \frac{1}{x'_y} = \frac{1}{\dfrac{1}{y \ln a}} = y \ln a = a^x \ln a$$

Hence,

$$(a^x)' = a^x \ln a$$

In particular $(e^x)' = e^x$.

10. *The Derivative of a Power Function.* According to the formula of the derivative of a composite function we have

$$(x^n)' = [(e^{\ln x})^n]' = (e^{n \ln x})' = e^{n \ln x} n \frac{1}{x} = x^n n \frac{1}{x} = nx^{n-1}$$

Thus,

$$(x^n)' = nx^{n-1}$$

This formula holds for any n, both integral and non-integral.

For example. $(\sqrt{x})' = (x^{\frac{1}{2}})' = \frac{1}{2} x^{\frac{1}{2}-1} = \frac{1}{2\sqrt{x}}$; $\left(\frac{1}{x}\right)' = (x^{-1})' =$

$= -1x^{-1-1} = -\frac{1}{x^2}$ and the like.

11. The Derivatives of Hyperbolic Functions. We have

$$(\sinh x)' = \left(\frac{e^x - e^{-x}}{2}\right)' = \frac{e^x - e^{-x}(-x)'}{2} = \frac{e^x + e^{-x}}{2} = \cosh x$$

Similarly,

$$(\cosh x)' = \sinh x;$$

$$(\tanh x)' = \left(\frac{\sinh x}{\cosh x}\right)' = \frac{(\sinh x)' \cosh x - \sinh x (\cosh x)'}{\cosh^2 x} =$$

$$= \frac{\cosh^2 x - \sinh^2 x}{\cosh^2 x} = \frac{1}{\cosh^2 x};$$

$$(\sinh^{-1} x)' = [\ln (x + \sqrt{x^2 + 1})]' =$$

$$= \frac{1}{x + \sqrt{x^2+1}} (x + \sqrt{x^2+1})' = \frac{1}{x + \sqrt{x^2+1}} \left(1 + \frac{(x^2+1)'}{2\sqrt{x^2+1}}\right) =$$

$$= \frac{1}{x + \sqrt{x^2+1}} \left(1 + \frac{2x}{2\sqrt{x^2+1}}\right) = \frac{1}{x + \sqrt{x^2+1}} \frac{\sqrt{x^2+1} + x}{\sqrt{x^2+1}} = \frac{1}{\sqrt{x^2+1}}!$$

These formulas also demonstrate a rather close analogy between trigonometric and hyperbolic functions.

12. The above formulas (comprising the table of basic differentiation* formulas) should be learnt by heart since they will be permanently used in what follows. With the help of the formulas it is possible to compute the derivative of any elementary function by using the rules of Sec. 4. For example,

$$(\sqrt[3]{x}\, 2^{\tan 5x})' = (\sqrt[3]{x})'\, 2^{\tan 5x} + \sqrt[3]{x}\, (2^{\tan 5x})'$$

But

$$(\sqrt[3]{x})' = (x^{\frac{1}{3}})' = \frac{1}{3} x^{\frac{1}{3}-1} = \frac{1}{3\sqrt[3]{x^2}}$$

and, by the formula of the derivative of a composite function, we obtain

$$(2^{\tan 5x})' = 2^{\tan 5x} \ln 2\, (\tan 5x)' = 2^{\tan 5x} \ln 2\, \frac{1}{\cos^2 5x} (5x)' =$$

$$= 2^{\tan 5x} \ln 2\, \frac{1}{\cos^2 5x} \cdot 5$$

---

* The operation of finding the derivative of a function is usually called differentiation (see Sec. 8).—*Tr.*

Taking the common factor outside the brackets we derive

$$\left(\sqrt[3]{x}\, 2^{\tan 5x}\right)' = \frac{2^{\tan 5x}}{3\sqrt[3]{x^2}}\left(1 + 15\ln 2 \,\frac{x}{\cos^2 5x}\right)$$

After some practice calculations of this type can be carried out much faster without intermediate transformations.

13. In some cases it is useful to take logarithms before calculating a derivative. For example, let it be necessary to find the derivative $(x^{\sin x})'$. Then we write $y = x^{\sin x}$; $\ln y = \sin x \ln x$ and $(\ln y)'_x = (\sin x \ln x)'$. Therefore,

$$\frac{1}{y}\, y' = \cos x \ln x + \sin x \,\frac{1}{x}$$

To calculate the left-hand side we have applied the formula of the derivative of a composite function. Finally, from the last relation,

$$y' = (x^{\sin x})' = x^{\sin x}\left(\cos x \ln x + \sin x \,\frac{1}{x}\right)$$

This method is sometimes used when it is necessary to find the derivative of the product of several factors since after taking the logarithm the product turns into a sum, and, generally speaking, it is easier to find the derivative of the sum than that of the product.

6. **Determining Tangent in Polar Coordinates.** The problem of determining the tangent to a curve which is represented by its equation in Cartesian coordinates was solved in Sec. 3. Now let a curve be given by its equation $\rho = f(\varphi)$ in polar coordinates. To determine the tangent we could transform the equation into Cartesian coordinates but it is simpler to solve the problem directly in polar coordinates. Let the position of the tangent be determined by the angle θ (see Fig. 114). Let us give φ a small increment $\Delta\varphi$ and let us consider the infinitesimal curvilinear triangle MNP formed by the two coordinate lines and the graph of $\rho = f(\varphi)$. For the sake of convenience the triangle is depicted in Fig. 114 separately. The triangle may be regarded as a "genuine" triangle (i.e. as a rectilinear triangle) to within to infinitesimals of higher order and even as a rectangular triangle since $\angle N = 90°$. (Why is it so?) This implies

$$\cot\theta = \lim_{\Delta\varphi\to 0}\cot\theta^* = \lim_{\Delta\varphi\to 0}\frac{\Delta\rho}{\rho\,\Delta\varphi} = \frac{1}{\rho}\,\rho'_\varphi \qquad (19)$$

Example 1. For the logarithmic spiral (see Sec. II.5) we have

$$\text{e of a cu}\quad \theta = \frac{1}{e^{k\varphi}}\,(ke^{k\varphi}) = k = \text{const}$$

Thus, the spiral intersects all the coordinate rays forming one and the same angle with them. It is easy to find the relationship between this property and the property indicated in Sec. II.5: they turn out to be equivalent.

Example 2. Let us consider the polar equation of a parabola with respect to its focus [i.e. equation (II.29) in which we should put
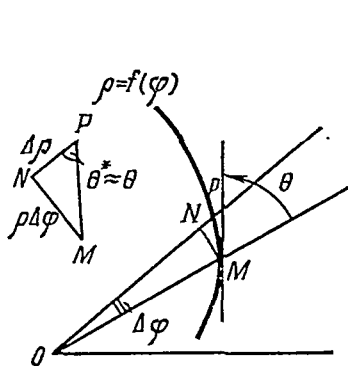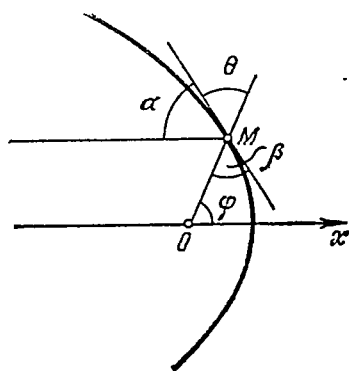
Fig. 114   Fig. 115

$\varepsilon = 1$]. By formula (19) we have

$$\cot \theta = \frac{1 + \cos \varphi}{p} \; \frac{p \sin \varphi}{(1 + \cos \varphi)^2} = \frac{\sin \varphi}{1 + \cos \varphi} = \frac{2 \sin \dfrac{\varphi}{2} \cos \dfrac{\varphi}{2}}{2 \cos^2 \dfrac{\varphi}{2}} =$$

$$= \tan \frac{\varphi}{2} = \cot \left(\frac{\pi}{2} - \frac{\varphi}{2} \right) ;$$

$$\theta = \frac{\pi}{2} - \frac{\varphi}{2}$$

Therefore if we draw a straight line parallel to the polar axis and passing through the point M we shall have (see Fig. 115) $\alpha + \theta = \pi - \varphi$, that is

$$\alpha = \pi - \varphi - \theta = \pi - \varphi - \left(\frac{\pi}{2} - \frac{\varphi}{2} \right) = \frac{\pi}{2} - \frac{\varphi}{2} = \theta = \beta$$

From this we obtain *the basic optical property of a parabola*: if the light propagates in the plane of the parabola from a light source placed in its focus then all the rays reflected by the parabola are
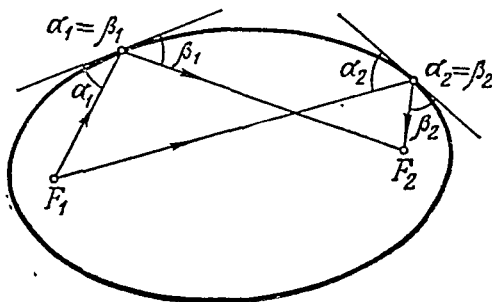
Fig. 116

parallel to its axis. Here lies the explanation of the fact that the reflector of a projector is shaped in the form of a surface generated by the revolution of a parabola about its axis (the parabolic reflector).

It is a little more complicated to deduce analogous optical properties of an ellipse and of a hyperbola. For instance, one can show that all the rays issued from a focus of an ellipse gather at the other focus after being reflected from the ellipse (see Fig. 116). A parabola may be regarded as an ellipse with one of its focuses removed to infinity (see the end of Sec. II.12) and therefore the optical property of a parabola is implied by that of an ellipse if we pass to the limit.

§ 2. *Differential*

7. Physical Examples. The notion of a differential is closely related to that of a derivative and is also one of the most important notions in mathematics. We shall illustrate it by considering the same examples as in Sec. 1.

Let a point have the velocity $v = s'_t = f'(t)$ at a moment t in its rectilinear motion according to the law of motion $s = f(t)$. If now some additional time Δt passes the point will cover some additional distance Δs. In case the motion is non-uniform the dependence of Δs on Δt can be complicated because the velocity of motion varies all the time. But if Δt (the time passed) is not large the velocity has no time to change considerably during the period of time from t to $t + \Delta t$. Therefore the motion may be regarded as "almost uniform" during this period. Hence, reckoning the distance we shall not get a serious error if we regard the motion as uniform, i.e. as having a constant velocity, namely, the velocity it had at the moment t.

Thus we obtain the distance $v \Delta t = s'_t \Delta t$. The distance is directly proportional to the time passed Δt. $s'_t \Delta t$ is called the **differential of the distance** and is denoted as ds: $ds = s'_t \Delta t$ (the symbol ds should be understood as an indivisible symbol and not as the product of d by s). The real distance Δs differs from the "invented" distance ds. of course, since the velocity may change during the period Δt no matter how small Δt is. But nevertheless if this period is small enough we can put approximately

$$\Delta s \approx ds \qquad (20)$$

But the smaller Δt, the smaller the change of the velocity. Therefore the accuracy of formula (20) becomes greater as Δt is decreased. In Sec. 8 we shall show that when the interval Δt is infinitesimal the difference between Δs and ds is an infinitesimal variable of higher order relative to Δs. There are many situations when it is permissible to neglect such infinitesimals of higher order. Then it is possible to say that the differential of the distance is nothing but an

infinitesimal distance, i.e. the distance corresponding to an infinitesimal interval of time. At the same time, of course, the differential of the distance may not be an infinitesimal at all in case Δt is not small, but the greater Δt, the lower the degree of accuracy of formula (20). Nevertheless, it is much easier to compute ds as a distance passed in a uniform motion than to evaluate the real distance Δs; this accounts for the fact that formula (20) is often used even when Δt is not very small.

Turning to the second example and reasoning in the same way we can say that the differential of the volume dV represents the volume which would be filled if the rate of filling remained constant and equal to the rate at the moment t during the period of time from t to $t + \Delta t$, that is $dV = V'_t \Delta t$. Similarly, the differential of the mass in the third example is the mass which the part AB of the curve (see Fig. 110) would have if the density of this part were constant and equal to the density at the point A, that is $dM = \rho \, \Delta s = M'_s \Delta s$.

In all cases the replacement of a real change of a quantity by its differential reduces to the transition from some non-uniform processes, non-homogeneous objects, etc. to the uniform and homogeneous ones. Such a replacement is always based upon the fact that every process is "almost uniform" during a small interval of time and every object is "almost homogeneous" in the small and so on.

8. Definition of Differential and Its Connection with Increment. Now we shall give the general definition of a differential. Let the argument of a function $y = f(x)$ first take a value x and then receive an increment Δx. Then the **differential of the function** is the product

$$dy = df(x) = y' \, \Delta x = f'(x) \, \Delta x \qquad (21)$$

The differential is therefore the increment which the function would receive if it changed in the interval from x to $x + \Delta x$ with the same velocity as for the value x of the argument.

The operation of finding the differential of a function is called differentiation of the function; it is carried out quite simply by means of formula (21). For example, let $y = \sin x$; then $dy = (\sin x)' \, \Delta x = \cos x \, \Delta x$, i.e. $d \sin x = \cos x \, \Delta x$.

Similarly, $d \tan x = \dfrac{1}{\cos^2 x} \, \Delta x$, $d(x^3) = 3x^2 \, \Delta x$ and the like. Thus, when differentiating a function one must find its derivative and multiply the result by Δx; therefore the operation of computing the derivative is also often called differentiation. But one must take care not to confuse the derivative with the differential. The derivative of a function $y = f(x)$ depends only on x whereas the differential also depends on Δx. In practical applications the differential is usually regarded as an infinitesimal whereas the derivative

is understood as a finite value. In case variables x and y have certain dimensions

$$[dy] = [\Delta y] = [y], \quad [y'_x] = \frac{[\Delta y]}{[\Delta x]} = \frac{[y]}{[x]}$$

We note, in particular, that

$$dx = x'_x \, \Delta x = 1 \, \Delta x = \Delta x$$

which means that *the differential of an independent variable is equal to its increment.* This makes it possible to rewrite formula (21):

$$dy = f'(x) \, dx = y' \, dx \tag{22}$$

and, on the other hand, to represent the derivative as the ratio of the differentials:

$$y'_x = \frac{dy}{dx} \quad \text{or, which is the same,} \quad f'(x) = \frac{df(x)}{dx}$$

The geometrical meaning of the differential of a function is shown in Fig. 117: *the differential is equal to the increment of the ordinate*
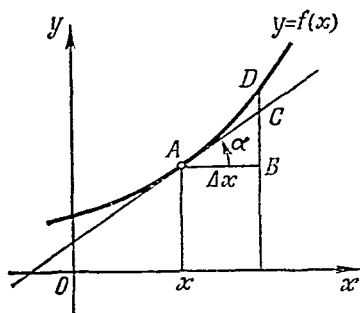
Fig. 117
$BD = \Delta y, \quad BC = AB \tan \alpha = \Delta x \cdot y' = dy$
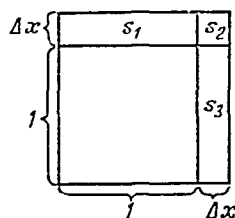
Fig. 118
$\Delta y = s_1 + s_2 + s_3, \quad dy = s_1 + s_3 \quad (\Delta x)^2 = s_2$

of the tangent. Hence, the replacement of the increment of a function by its differential is equivalent to the replacement of the graph of the function by the segment of the tangent drawn through the point A. This replacement is justified in case Δx is small enough.

To investigate the connection between the differential and the increment we take into account that $\frac{\Delta y}{\Delta x} \xrightarrow[\Delta x \to 0]{} y'$, i.e. $\frac{\Delta y}{\Delta x} = y' + \alpha$ where $\alpha \to 0$ as $\Delta x \to 0$. This yields

$$\Delta y = y' \, \Delta x + \alpha \, \Delta x = dy + \beta \tag{23}$$

where $\beta = \alpha \, \Delta x$ is an infinitesimal variable of higher order than Δx (it is represented by the segment CD in Fig. 117). The equality

(23) can be formulated as *the differential is the principal linear part of the increment of a function*. It is called here the principal part since the difference between the differential and the increment is the infinitesimal β of higher order and it is called the linear part since it is directly proportional to Δx (compare with Sec. I.22). If $y' \neq 0$ then dy and $dx = \Delta x$ are infinitesimals of the same order and therefore β in formula (23) is an infinitesimal of higher order than dy, that is dy and Δy are equivalent infinitesimals (see Sec. III.8).

Let us take an example to illustrate the error which occurs if the increment of a function is replaced by its differential. Let $y = x^2$ and let the argument first assume the value $x = 1$ and then receive the increment Δx. We have

$$\Delta y = (1 + \Delta x)^2 - 1^2 = 2\,\Delta x + \Delta x^2;$$
$$dy = y'\,\Delta x = 2 \cdot 1 \cdot \Delta x = 2\,\Delta x$$

Therefore, Δy and dy differ from each other by the infinitesimal $(\Delta x)^2$ of the second order (see Fig. 118). In particular,

for $\Delta x =$	0.1	$\Delta y =$ 0.21;	$dy =$ 0.2;	the error is 5 per cent;
for $\Delta x =$	0.01	$\Delta y =$ 0.0201;	$dy =$ 0.02;	the error is 0,5 per cent;
for $\Delta x =$	-0.001	$\Delta y = -0.001999$;	$dy = -0.002$;	the error is 0.05 per cent etc.

It is obvious here that the relative error generated by the replacement of Δy by dy decreases rapidly as $|\Delta x|$ decreases.

A function which has the differential is called **differentiable**. In other words, a differentiable function is a function such that its small increment has the principal linear part, i.e. a function which may be approximately replaced by a linear function on every small interval of change of the argument (such a replacement is the so-called process of **linearizing**). A differentiable function must have a finite derivative, and the function itself must be continuous for the considered values of the argument since (23) shows that Δy is infinitesimal when Δx is infinitesimal. At the same time a continuous function may turn out to be non-differentiable at some points. For instance, the function shown in Fig. 113 is non-differentiable not only at the point of discontinuity $x = x_3$ but also at the points $x = x_1$ and $x = x_2$ where it is continuous. B. Bolzano, a famous Czech mathematician (1781-1848), and, independently of him, K. Weierstrass (1815-1897), a prominent German mathematician, discovered (the former in 1830 and the latter in 1860; Bolzano's result was not published) the existence of continuous functions which are non-differentiable for all values of the argument. Such functions had been considered a mathematical trick for a long time but it

turned out that they were of essential importance for describing processes of the type of a Brownian motion. We shall not pay attention to the possibility of existence of such "monsters" in our introductory course.

9. Properties of Differential. The differential of a function is obtained by multiplying its derivative by the differential of the argument and therefore each property of the derivative (see Sec. 4) obviously implies the corresponding property of the differential. For example, multiplying both parts of the equality $(u + v)' = u' + v'$ by dx we receive $(u + v)' dx = u' dx + v' dx$ or, which is the same,

$$d(u + v) = du + dv$$

(i.e. *the differential of a sum is equal to the sum of the differentials*). Similarly, we deduce the formula

$$d(uv) = (du) v + u dv \tag{24}$$

and the like. We shall see in Sec. IX.12 that these formulas also hold for the case of an arbitrary number of independent variables.

The implication of the formula for the derivative of a composite function is of special importance. Let $y = f(x)$ and let x first be an independent variable. Then each of the formulas (21) and (22) can be used for calculating dy since in this case $\Delta x = dx$. Now let x depend on a third variable, for example, $x = x(t)$. Then $\Delta x \neq dx$ but it turns out that nevertheless formula (22) remains true [whereas formula (21) does not hold, in general]. Virtually,

$$dy = y'_t dt = y'_x x'_t dt = y'_x dx$$

which is what we set out to prove. Therefore it is natural to use formula (22) [and not (21)] for calculating the differential since this formula remains true (invariant) in all cases.

Now we shall apply this invariance property* to computing the derivative of a function represented parametrically (see Sec. II.6). Let $x = x(t)$ and $y = y(t)$ (t is a parameter). Then $dx = \dot{x} dt$ and $dy = \dot{y} dt$ (the dot usually denotes the derivative with respect to a parameter) which implies

$$y' = \frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} \tag{25}$$

All the properties of differentials are used, in particular, for linearizing relations between variables, that is for passing from a general, non-linear, relation to the linear relation between the increments of the variables. Such a linearization is possible in case

---

* This property is usually called the invariance of the form of the differential.—*Tr.*

the changes of the variables are small, and it is based on dropping infinitesimals of higher order.

Thus, for instance, equation (II.30) characterizes the non-linear relation between the coordinates of a point $M(x, y)$ belonging to a curve of the second order. But now let the position of the point M change near a fixed point $M_0(x_0, y_0)$, that is let the increments $x - x_0 = \xi$ and $y - y_0 = \eta$ be small. Differentiating equation (II.30) and replacing the differentials by the increments we come to the linearized equation

$$2Ax_0\xi + 2By_0\xi + 2Bx_0\eta + 2Cy_0\eta + D\xi + E\eta = 0 \qquad (26)$$

which describes the approximate linear relation between ξ and η. Since when deducing equation (26) we replaced the differentials dx and dy by the increments ξ and η, the point M of the line satisfies the equation only with an accuracy of the infinitesimals of higher order. Equation (26) is precisely satisfied by the points of the tangent line to curve (II.30) drawn through the point M_0.

The linearization is widely used in physics, in particular, when differential equations are deduced (see Sec. XIV.6).

10. Application of Differentials to Approximate Calculations. Differentials are widely used for approximate calculations. First of all, the increment of a function is often replaced by its differential which, as a rule, can be found in a simpler way.

Suppose we are given a function $y = f(x)$. Let a particular value $f(a)$ be known. Let the argument x receive a small increment $\Delta x = h$. Then we can put

$$f(a + h) - f(a) = \Delta y \approx dy = f'(a)h$$

that is

$$f(a + h) \approx f(a) + f'(a)h \qquad (27)$$

Choosing the concrete functions $\sqrt[n]{x}$, $\sin x$, $\ln x$ and so forth as $f(x)$ we derive the approximate formulas

$$\left.\begin{array}{c} \sqrt[n]{a+h} \approx \sqrt[n]{a} + \dfrac{h}{n\sqrt[n]{a^{n-1}}} \\[2ex] \sin(a+h) \approx \sin a + h\cos a \\[2ex] \ln(a+h) \approx \ln a + \dfrac{h}{a} \end{array}\right\} \qquad (28)$$

etc. which are applicable for small $|h|$. The formulas can be specified and their errors can be effectively estimated. This question will be discussed in Secs. 15-16.

Let us consider an example. Suppose we know that $\ln 2 = 0.693$. Then calculating with an accuracy of 0.001 we get

$$\ln 2.1 = \ln(2 + 0.1) \approx \ln 2 + \frac{0.1}{2} = 0.693 + 0.050 = 0.743$$

The table of logarithms gives the value $\ln 2.1 = 0.742$, i.e. the error is smaller than 0.2 per cent.

It is sometimes necessary to transform the expression that must be calculated in order to facilitate the calculations. For instance, it is wrong to calculate $\sqrt[3]{2}$ as

$$\sqrt[3]{2} = \sqrt[3]{1+1} \approx \sqrt[3]{1} + \frac{1}{3\sqrt[3]{1^2}} = 1 + \frac{1}{3} = 1.333$$

since the value $h = 1$ can hardly be regarded as small in comparison with $a = 1$. It is convenient to put here $\sqrt[3]{2} = \frac{\sqrt[3]{2m^3}}{m}$ and to choose the integer m so that $2m^3$ should become as close as possible to an exact cube of an integer. It is possible to take $m = 4$ since $2 \cdot 4^3 = 128$ is close to $125 = 5^3$; then we get

$$\sqrt[3]{2} = \frac{1}{4} \sqrt[3]{2 \cdot 4^3} = \frac{1}{4} \sqrt[3]{128} = \frac{1}{4} \sqrt[3]{125+3} \approx$$

$$\approx \frac{1}{4} \left(\sqrt[3]{125} + \frac{3}{3\sqrt[3]{125^2}} \right) = \frac{1}{4} \left(5 + \frac{1}{25} \right) = \frac{1}{4} \times 5.0400 = 1.2600$$

Tables of roots yield the value $\sqrt[3]{2} = 1.2599$, i.e. the error is smaller than 0.01 per cent.

Differentials are also used for estimating errors. Suppose, the variables x and y are connected by a functional relation $y = f(x)$, and let the approximate value $\overline{x}$ of the argument x be known with the maximum absolute error α_x (see Sec. I.7). Then, of course, $\overline{y} = f(\overline{x})$ should be taken as an approximate value of y. To estimate the maximum absolute error α_y we observe that $x = \overline{x} + h$ where $|h| < \alpha_x$ and therefore, if α_x (and, consequently, h) is small, then

$$y = \overline{y} + \Delta y \approx \overline{y} + dy = \overline{y} + f'(\overline{x}) h$$

that is

$$|y - \overline{y}| \approx |f'(\overline{x})| \cdot |h| < |f'(\overline{x})| \alpha_x$$

Thus, we can put

$$\alpha_y = |f'(\overline{x})| \alpha_x \qquad (29)$$

For example, let $y = x^n$. Then

$$\alpha_y = |n\overline{x}^{n-1}| \alpha_x$$

and the corresponding maximum relative errors are connected by the simple formula

$$\delta_y = \frac{\alpha_y}{|\overline{y}|} = \frac{|n\overline{x}^{n-1}| \alpha_x}{|\overline{x}|^n} = \frac{|n| \alpha_x}{|\overline{x}|} = |n| \delta_x$$

As another example let us consider $\ln 10.7$ where the value 10.7 is approximate and is known with an accuracy of 0.1. We have $\ln 10.7 = 2.3702$ according to the table but it is obvious that this result contains too many decimal digits. To understand what the accuracy of the result is we must take into account that in our case $\alpha_x = 0.1$ which implies

$$\alpha_y = \frac{1}{10.7} 0.1 \approx 0.01$$

that is the result should be put down in the form $\ln 10.7 = 2.37$.

§ 3. Derivatives and Differentials of Higher Orders

11. Derivatives of Higher Orders. Let $y = f(x)$. Then the derivative $y' = f'(x)$ which was studied in § 1 is called the **derivative of the first order** or the **first derivative** of the function $f(x)$. In its turn, $f'(x)$ is also a function of x and therefore it is possible to take its derivative which is called the **derivative of the second order** or the **second derivative** of the original function:

$$y'' = (y')' = f''(x)$$

In the same way we define the **derivative of the third order** (the **third derivative**):

$$y''' = (y'')' = f'''(x)$$

The consequent derivatives are denoted as $y^{(4)} = y^{\mathrm{IV}}$, $y^{(5)} = y^{\mathrm{V}}$ etc. For example, $(x^3)' = 3x^2$, $(x^3)'' = (3x^2)' = 6x$; $(\sin x)' = \cos x$, $(\sin x)'' = (\cos x)' = -\sin x$ and the like. The derivative of the second order sometimes has a clear physical meaning: thus, in the first example of Sec. 1 the derivative of the second order of the distance with respect to the time is the velocity of change of the instantaneous velocity, that is the instantaneous acceleration. We shall discuss the applications of the derivatives of higher orders in Sec. 15 and further.

The formula for the derivative of a sum is quite simple. If $y = u + v$ then $y' = u' + v'$, $y'' = (u' + v')' = u'' + v''$ and so on. Generally,

$$(u + v)^{(n)} = u^{(n)} + v^{(n)}$$

As for the formula for the derivative of a product, we have
$(uv)' = u'v + uv'$,

$(uv)'' = (u'v + uv')' = u''v + u'v' + u'v' + uv'' =$
$\quad = u''v + 2u'v' + uv''$,

$(uv)''' = (u''v + 2u'v' + uv'')' = u'''v + 2u''v' + u'v'' + u''v' +$
$\quad + 2u'v'' + uv''' = u'''v + 3u''v' + 3u'v'' + uv'''$ etc.   (30)

Here computing the next derivative we begin with differentiating the first factors in all the terms and after that we differentiate the second factors in all the terms. These computations are carried out according to the scheme resembling that of expanding the expressions $(a + b)^2$, $(a + b)^3$ and so forth:

$$(a + b)^2 = (a + b)(a + b) = a^2 + ab + ab + b^2 =$$
$$= a^2 + 2ab + b^2,$$
$$(a + b)^3 = (a^2 + 2ab + b^2)(a + b) = a^3 + 2a^2b + ab^2 +$$
$$+ a^2b + 2ab^2 + b^3 = a^3 + 3a^2b + 3ab^2 + b^3 \text{ etc.} \quad (31)$$

Therefore the coefficients in formulas (30) are the same as in formulas (31). In the general case the formula (the Leibniz rule) can be written as

$$(uv)^{(n)} = u^{(n)}v + \binom{n}{1} u^{(n-1)} v' + \binom{n}{2} u^{(n-2)} v'' + \ldots + uv^{(n)} \quad (32)$$

where $\binom{n}{1}$, $\binom{n}{2}$, ... are the so-called **binomial coefficients**, i.e. the coefficients occurring in the binomial formula [the expansion of the power $(a + b)^n$].

The calculation of the derivative of an implicit function will be illustrated by taking equation (12). Further differentiation of equality (13) yields

$$\frac{2}{a^2} + \frac{2}{b^2} (y'y' + yy'') = 0, \quad \text{i.e.} \quad y'' = -\frac{b^2}{y} \left(\frac{1}{a^2} + \frac{y'^2}{b^2} \right) =$$
$$= -\frac{b^2}{y} \left(\frac{1}{a^2} + \frac{b^2x^2}{a^4y^2} \right) = -\frac{b^4}{a^2y^3} \left(\frac{y^2}{b^2} + \frac{x^2}{a^2} \right) = -\frac{b^4}{a^2y^3} \quad (33)$$

[here we have used expression (14) for y']. The following derivatives are computed in a similar way. Formula (33) can also be obtained by differentiating equality (14).

In addition, we shall turn to differentiating a function represented by its parametric equations. Differentiating formula (25) we get

$$y'' = \frac{d(y')}{dx} = \frac{d\left(\dfrac{\dot{y}}{\dot{x}} \right)}{dx} = \frac{\dfrac{\ddot{y}\dot{x} - \dot{y}\ddot{x}}{\dot{x}^2} dt}{\dot{x} \, dt} = \frac{\ddot{y}\,\dot{x} - \dot{y}\,\ddot{x}}{\dot{x}^3}$$

(the two dots denote the second derivative with respect to a parameter). Here, as in deducing formula (25), we need not pay attention to the fact which of the variables is independent when we calculate the differential (the first differential, as we shall call it in Sec. 12). The derivatives of higher order can be found similarly.

12. Higher-Order Differentials. Let $y = f(x)$. Then

$$dy = f'(x) \, dx \quad (34)$$

The differential calculated by formula (34) is called the first differential (the first-order differential). The differential of the second order (the second-order differential) of a function is the first differential of the first differential of the function. It is denoted by d^2y. If x is an independent variable then in the process of the repeated differentiation the quantity dx is regarded as being independent of x. Thus, dx being a constant, we take it outside the differentiation sign:

$$d^2y = d\,(dy) = d\,(f'\,(x)\,dx) = d\,(f'\,(x))\,dx =$$
$$= (f'\,(x))'\,dx\,dx = f''\,(x)\,dx^2 \tag{35}$$

where the notation $dx^2 = (dx)^2$ is assumed. Similarly, we obtain

$$d^3y = d\,(d^2y) = f'''\,(x)\,dx^3 \tag{36}$$

and so on. This enables us to write the derivatives of higher order as the ratios of the corresponding differentials:

$$y'' = \frac{d^2y}{dx^2}\,, \qquad y''' = \frac{d^3y}{dx^3} \quad \text{etc.} \tag{37}$$

In addition, we see that if dy is an infinitesimal of the first order relative to dx then d^2y is an infinitesimal of the second order, d^3y is an infinitesimal of the third order and so on. Further, we note that

$$d^2x = d\,(dx) = d\,(1 \cdot dx) = dx\,d\,(1) = 0$$

that is *the second-order differential of an independent variable is equal to zero*; of course, all the following, differentials of an independent variable are also equal to zero.

In case x is not an independent variable (or we do not know whether it is independent or not) formula (34), as it was seen in Sec. 8, is nevertheless true. But using this formula for subsequent differentiations we must not consider dx constant but should use the rule of differentiating a product [see formula (24)]:

$$d^2y = d\,(f'\,(x)\,dx) = d\,(f'\,(x)) \cdot dx + f'\,(x)\,d\,(dx) =$$
$$= f''\,(x)\,dx^2 + f'\,(x)\,d^2x \tag{38}$$

In a similar way we find

$$d^3y = f'''\,(x)\,dx^3 + 3f''\,(x)\,dx\,d^2x + f'\,(x)\,d^3x \tag{39}$$

(check it!) and all the following differentials. If now it turns out that x is an independent variable then $d^2x = d^3x = 0$, and formula (38) turns into formula (35) and formula (39) turns into formula (36). Thus, *formulas (35)-(37) should be used only in case x is an independent variable*.

The notion of the derivative and that of the differential and all the basic rules of operating on them were elaborated by Newton

(1666) and by Leibniz (1684) although these notions had been used for some particular problems before.

Differential calculus has very many applications in the field of investigating the behaviour of functions. These applications will be considered in the forthcoming sections of our course.

§ 4. L'Hospital's Rule

13. **Indeterminate Forms of the Type** $\frac{0}{0}$. We said in Sec. III.7 that the evaluation of the limit of the ratio of two infinitesimals might yield different results in different cases. J. Bernoulli discovered a simple rule for evaluating such a limit which is applicable to many cases. This rule was published in 1696 in the first printed textbook on differential calculus written by L'Hospital, a French mathematician (1661-1704). Let it be necessary to evaluate the limit

$$\lim_{t \to t_0} \frac{\varphi(t)}{\psi(t)} \tag{40}$$

and

$$\varphi(t_0) = \psi(t_0) = 0 \tag{41}$$

that is let us have an indeterminate form of the type $\frac{0}{0}$. Suppose that the limit (finite or infinite)

$$\lim_{t \to t_0} \frac{\varphi'(t)}{\psi'(t)} = k \tag{42}$$

is found in some way. Then we assert that the limit (40) is also equal to k, that is we have

$$\lim_{t \to t_0} \frac{\varphi(t)}{\psi(t)} = \lim_{t \to t_0} \frac{\varphi'(t)}{\psi'(t)} \tag{43}$$

for indeterminate forms of the type $\frac{0}{0}$.

In order to prove the rule let us consider the curve $y = \varphi(t)$, $x = \psi(t)$ in the x, y-plane. Then, as $t \to t_0$, by (41), this curve approaches the origin of the coordinate system. To find out in what way it approaches the origin (i.e. like a spiral or along a certain direction which then should be specified etc.) we note that, by (42), we have

$$\frac{dy}{dx} = \frac{\varphi'(t)\,dt}{\psi'(t)\,dt} = \frac{\varphi'(t)}{\psi'(t)} \longrightarrow k \quad (\varepsilon s \ t \to t_0)$$

Hence (see Fig. 119), while the curve approaches the origin the tangent to the curve turns in this process and tends to the limiting position in which it forms an angle α_0 with the x-axis such that $\tan \alpha_0 = k$. But then the angle β ("the angle of elevation") also

tends to α_0, i.e.

$$\frac{\varphi(t)}{\psi(t)} = \frac{y}{x} = \tan \beta \xrightarrow[t \to t_0]{} \tan \alpha_0 = k \tag{44}$$

which is just what we set out to prove.

It sometimes happens that using L'Hospital's rule we obtain a ratio of derivatives which is again an indeterminate form of the type $\frac{0}{0}$; then it is possible to apply the rule again and so forth.

For example,

$$\lim_{x \to 0} \frac{x - \sin x}{x^3} = \left(\frac{0}{0}\right) = \lim_{x \to 0} \frac{1 - \cos x}{3x^2} = \left(\frac{0}{0}\right) = \lim_{x \to 0} \frac{\sin x}{6x} = \left(\frac{0}{0}\right) =$$

$$= \lim_{x \to 0} \frac{\cos x}{6} = \frac{1}{6},$$

$$\lim_{x \to 1} \frac{2^x - 4 \times 2^{-x}}{(x-1)^2} = \left(\frac{0}{0}\right) = \lim_{x \to 1} \frac{2^x \ln 2 + 4 \times 2^{-x} \ln 2}{2(x-1)} = \frac{4 \ln 2}{0} = \pm \infty$$

We have applied L'Hospital's rule three times in the first example and once in the second example.

L'Hospital's rule for indeterminate forms of type (40) always achieves the aim in case t_0 is a finite number and the numerator
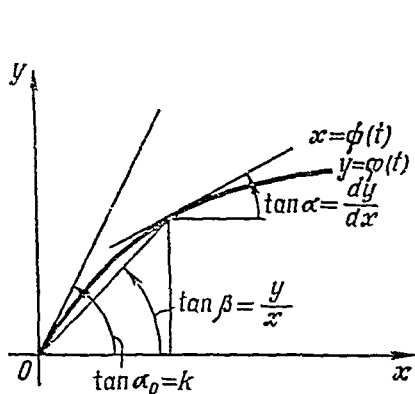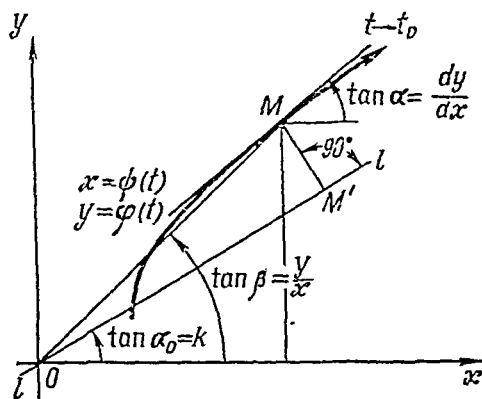
Fig. 119   Fig. 120

and the denominator are infinitesimals of an integral order of smallness relative to $t - t_0$ (see Sec. III.10). Indeed, L'Hospital's rule implies that each differentiation reduces the order of the infinitesimals by unity, and after several steps we shall obtain variables of the "zero order" (that is having a finite limit) in the numerator or in the denominator (or in both of them) and thus we shall no longer have an indeterminate form.

14. Indeterminate Forms of the Type $\frac{\infty}{\infty}$. L'Hospital's rule (43) also remains true for indeterminate forms of the type $\frac{\infty}{\infty}$, i.e. when the condition

$$| \varphi (t_0) | = | \psi (t_0) | = \infty$$

is put instead of (41).

The proof for this case is analogous to the one given in Sec. 13 but now the curve $y = \varphi (t)$, $x = \psi (t)$ does not approach the origin of the coordinate system as $t \to t_0$ but travels to infinity (see Fig. 120). In this process the curve, by condition (42), turns in such a way that the angle α which the curve (i.e. its tangent) forms with the x-axis tends to α_0 where $\tan \alpha_0 = k$. But then the distance passed by the point M of the curve along the straight line ll (see Fig. 120) will be an infinitely large variable of higher order than the distance along the straight line transversal to ll, namely, $MM' \ll OM$. Hence, $\angle MOM' \to 0$ as the point M travels to infinity and therefore the "angle of elevation" β tends to α_0 and we can write the same formula (44) again which concludes the proof.

We give here several important examples:

$$\lim_{x \to \infty} \frac{\ln x}{x^k} = \left(\frac{\infty}{\infty} \right) = \lim_{x \to \infty} \frac{\frac{1}{x}}{kx^{h-1}} = \lim_{x \to \infty} \frac{1}{kx^k} = \frac{1}{\infty} = 0 \quad (k > 0),$$

$$\lim_{x \to \infty} \frac{x}{a^x} = \left(\frac{\infty}{\infty} \right) = \lim_{x \to \infty} \frac{1}{a^x \ln a} = \frac{1}{\infty} = 0 \quad (a > 1),$$

$$\lim_{x \to \infty} \frac{x^k}{b^x} = \lim_{x \to \infty} \left(\frac{x}{(\sqrt[k]{b})^x} \right)^k = \lim_{x \to \infty} \left(\frac{x}{a^x} \right)^k = 0^k = 0$$

$$(k > 0, \ b > 1, \ a = \sqrt[k]{b}).$$

Hence, a logarithmic function with a base greater than unity tends to infinity (as its argument tends to infinity) slower than any power function with a positive exponent. Besides, a power function tends to infinity slower than any exponential function with a base greater than unity.

L'Hospital's rule can be applied to some indeterminate forms of other types (see Sec. III.5, property 6 and the beginning of Sec. III.15) after they are transformed into forms of the type $\frac{0}{0}$ or $\frac{\infty}{\infty}$. This can be achieved according to the following scheme:

$$0 \cdot \infty = 0 \frac{1}{0} = \frac{0}{0}, \quad \infty - \infty = \frac{1}{0} - \frac{1}{0} \approx \frac{0-0}{0 \times 0} = \frac{0}{0}$$

These formulas should be understood in a conditional sense. We use them only to indicate the types of the variables. After taking logarithms we can also apply L'Hospital's rule to power indeterminate forms.

§ 5. Taylor's Formula and Series

15. Taylor's Formula. It was shown in Sec. 10 that replacing the increment of a function by its differential we can deduce many approximate formulas. It turns out that these formulas can be made much more accurate if we apply differentials of higher order. This problem is solved by means of Taylor's formula named after B. Taylor (1685-1731), an English mathematician.

Let us first suppose that we are given a polynomial $P(x)$. A polynomial is usually considered as expanded into powers of x but it is quite easy to expand it in powers of $x - a$ where a is an arbitrary number.

Suppose, for example, that we are going to expand the polynomial $P(x) = 5 - 3x + 2x^3$ in powers of $x - 4$. In order to do this it is sufficient to substitute $x = [4 + (x - 4)]$ and then remove the square brackets without removing the parentheses:

$$P(x) = 5 - 3[4 + (x - 4)] + 2[4 + (x - 4)]^3 =$$
$$= 5 - 12 - 3(x - 4) + 128 + 96(x - 4) + 24(x - 4)^2 +$$
$$+ 2(x - 4)^3 = 121 + 93(x - 4) + 24(x - 4)^2 + 2(x - 4)^3$$

In the general case, for a polynomial of degree n, we can write

$$P(x) = a_0 + a_1(x - a) + a_2(x - a)^2 +$$
$$+ a_3(x - a)^3 + \ldots + a_n(x - a)^n \qquad (45)$$

The coefficients here can be found in the following way. First we put $x = a$ and obtain $P(a) = a_0$. Then we differentiate formula (45):

$$P'(x) = a_1 + 2a_2(x - a) + 3a_3(x - a)^2 + \ldots$$
$$\ldots + na_n(x - a)^{n-1}$$

If now we put here $x = a$ this will yield $P'(a) = a_1$. Let us differentiate once again:

$$P''(x) = 1 \times 2a_2 + 2 \times 3a_3(x - a) + \ldots$$
$$\ldots + (n - 1)na_n(x - a)^{n-2}$$

which implies $P''(a) = 1 \quad 2a_2$. Further, in a similar way we derive $P'''(a) = 1 \times 2 \times 3a_3$ and so on. Generally, $P^{(k)}(a) = k!a_k$ (where $k! = \times 1\, 2 \ldots k$) which yields

$$a_k = P^{(k)}(a)/k!$$

Thus, formula (45) can be rewritten in the following form:

$$P(x) = P(a) + \frac{P'(a)}{1!}(x - a) + \frac{P''(a)}{2!}(x - a)^2 + \ldots$$

$$\ldots + \frac{P^{(n)}(a)}{n!}(x - a)^n = P(a) + \sum_{k=1}^{n} \frac{P^{(k)}(a)}{k!}(x - a)^k \qquad (46)$$

where $\sum$ is the summation sign (see Sec. III.6).

For example, taking the polynomial from the previous example we deduce

$$P'(x) = -3 + 6x^2, \quad P''(x) = 12x, \quad P'''(x) = 12,$$

$$P(4) = 5 - 3 \times 4 + 2 \times 4^3 = 121, \quad \frac{P'(4)}{1!} = -3 + 6 \times 4^2 = 93.$$

$$\frac{P''(4)}{2!} = \frac{12 \times 4}{2} = 24. \quad \frac{P'''(4)}{3!} = \frac{12}{6} = 2$$

that is we obtain the same values of the coefficients as before.

If now we take an arbitrary function $f(x)$ in place of a polynomial $P(x)$ then formula (46) will no longer hold. But if we denote the difference between the left-hand and the right-hand sides of formula (46) by $R_n(x)$ (the **remainder**) then we can write

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

$$\cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n(x) \tag{47}$$

It is this formula that is called **Taylor's formula.** The most essential thing about the formula is that *the remainder is an infinitesimal of at least $(n+1)$th order relative to $x - a$ as $x \to a$*, that is $R_n(x)$ is an infinitesimal of higher order than the last of the "exact" terms put down in formula (47). In order to prove this assertion let us suppose, for simplicity's sake, that $n = 2$. that is

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + R_2(x)$$

Finding $R_2(x)$ from this and applying L'Hospital's rule (see Sec. 13) we receive

$$\lim_{x \to a} \frac{R_2(x)}{(x-a)^3} = \lim_{x \to a} \frac{f(x) - f(a) - f'(a)(x-a) - \frac{f''(a)}{2!}(x-a)^2}{(x-a)^3} =$$

$$= \left(\frac{0}{0}\right) = \lim_{x \to a} \frac{f'(x) - f'(a) - f''(a)(x-a)}{3(x-a)^2} = \left(\frac{0}{0}\right) =$$

$$= \lim_{x \to a} \frac{f''(x) - f''(a)}{3 \times 2(x-a)} = \left(\frac{0}{0}\right) = \lim_{x \to a} \frac{f'''(x)}{3!} = \frac{f'''(a)}{3!} \tag{48}$$

i.e. the ratio $\frac{R_2(x)}{(x-a)^3}$ has a finite limit as $x \to a$. This just implies (see Sec. III.10) our assertion about the order of smallness of $R_2(x)$, and an analogous consideration holds for $R_n(x)$.

Let us denote $x = a + h$. We see that dropping the remainder in formula (47) for successively increasing values of n we shall obtain approximate formulas with, respectively, increasing degrees

of accuracy (for small values of $|h|$):

$$f(a+h) \approx f(a) + f'(a)\,h \qquad (49)$$

[this formula coincides with formula (27) and guarantees an accuracy to a term of the order of h^2],

$$f(a+h) \approx f(a) + f'(a)\,h + \frac{f''(a)}{2}\,h^2 \qquad (50)$$

with an accuracy of the order of h^3,

$$f(a+h) \approx f(a) + f'(a)\,h + \frac{f''(a)}{2}\,h^2 + \frac{f'''(a)}{6}\,h^3 \qquad (51)$$

with an accuracy of the order of h^4 and so forth.

The polynomials (in $h = x - a$) entering into the right-hand sides are called **Taylor's polynomials**. They yield, in some sense, the best approximate expression of a function $f(x)$ in the form of a polynomial of a given degree near the value $x = a$. Namely, a Taylor polynomial differs from $f(x)$ by a term which is an infinitesimal of the highest order in comparison with all the polynomials of the same degree as $x \to a$. For instance, even if we change only one of the coefficients on the right-hand side of (50) then the difference may become an infinitesimal of order 0, 1 or 2 but not an infinitesimal of the third order relative to $x - a$ (as $x \to a$).

16. Taylor's Series. Since the errors of formulas (49), (50), (51) etc. are becoming infinitesimals of greater and still greater order (as $n \to \infty$) it is quite natural to expect that for small $|h|$ it is permissible to pass to the limit and thus obtain an "exact" representation of $f(a+h)$ in the form of a sum of an infinite series (see Sec. III.6), that is

$$f(a+h) = f(a) + \frac{f'(a)}{1!}\,h + \frac{f''(a)}{2!}\,h^2 + \ldots + \frac{f^{(n)}(a)}{n!}\,h^n + \ldots$$

$$\ldots = f(a) + \sum_{n=1}^{\infty} \frac{f^{(n)}(a)}{n!}\,h^n \qquad (52)$$

This series is called **Taylor's series**; it was originally introduced by B. Taylor in 1715. Such series will be systematically treated in § 3 of Chapter XVII. It will be shown there that the above-mentioned supposition is true. By the way, we shall answer the question what particular values of h guarantee the possibility of using formula (52). It turns out that it is always permissible to apply the formula if the series is practically convergent in the sense described in the end of Sec. III.6 [but in this case the function which should be expanded into Taylor's series must not be represented by different formulas on different parts of the range of its argument (see Sec. I.13)]. Taking this comment into account we shall now use Taylor's series.

Formula (52) can be rewritten (by substituting $a + h = x$ and $h = x - a$) in the form

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots \qquad (53)$$

which is an expansion into powers of $x - a$. In particular, when $a = 0$ we obtain an expansion in powers of x:

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \ldots \qquad (54)$$

Series (54) is sometimes called Maclaurin's series after C. Maclaurin (1698-1746), a Scotch mathematician, which is historically incorrect.

For example, let $f(x) = e^x$. Then $f'(x) = e^x$, $f''(x) = e^x$, ... and

$$f(0) = 1, \quad f'(0) = 1, \quad f''(0) = 1, \quad \ldots$$

Hence formula (54) results here in

$$e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \ldots \qquad (55)$$

Let us evaluate the number e with an accuracy of 0.001. In order to do this let us put $x = 1$ and evaluate the terms one after another retaining a reserve decimal digit:

$$e = 1.0000 + 1.0000 + 0.5000 + 0.1667 + 0.0417 +$$
$$+ 0.0083 + 0.0014 + 0.0002 + 0.0000$$

Every subsequent term here is obtained by dividing the foregoing term by the next integer. As it is seen, the terms of the series manifest an obvious tendency to decrease fast and, in addition, they soon become less than the required degree of accuracy. Now summing up and rounding the result we obtain $e = 2.718$.

In a way similar to that of deducing (55) it is possible to receive the following formulas (we leave this to the reader):

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots \qquad (56)$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots \qquad (57)$$

$$\cosh x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \ldots \qquad (58)$$

$$\sinh x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \ldots \qquad (59)$$

$$(1+x)^p = 1 + \binom{p}{1}x + \binom{p}{2}x^2 + \binom{p}{3}x^3 + \ldots \qquad (60)$$

(for any arbitrary p) where the binomial coefficients $\binom{p}{k}$ are defined by the formulas $\binom{p}{1} = p$, $\binom{p}{2} = \frac{p(p-1)}{1 \cdot 2}$, $\ldots$, $\binom{p}{k} =$ $= \frac{p(p-1) \ldots (p-k+1)}{k!}$. If p is a positive integer, $\binom{p}{p+1} =$ $= \binom{p}{p+2} = \ldots = 0$ and series (60) turns into a finite sum. We thus obtain the expression for binomial coefficient.

For the, logarithmic function we similarly derive

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \tag{61}$$

J. L. Lagrange (1736–1813), a prominent French mathematician, proved that the remainder $R_n(x)$ in formula (47) admits the representation

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1}$$

where ξ is a certain value lying between a and x. This representation sometimes makes it possible to find out for what values of x formula (53) holds because the formula is true if and only if $R_n(x) \rightarrow$ $\rightarrow 0$ as $n \rightarrow \infty$. For instance, if we take into account that $\frac{h^n}{n!} \xrightarrow[n\to\infty]{} 0$ for any h it is easy to prove that formulas (55)–(59) are true for any x (think it over!).

Taylor's series can be rewritten in another form if we denote $x - a = \Delta x$; $f(x) - f(a) = \Delta f$; $f'(a)(x-a) = f'(a)\Delta x = df$; $f''(a)(x-a)^2 = f''(a)(\Delta x)^2 = d^2 f$ (see Sec. 12) etc. Then we deduce from (53),

$$\Delta f = df + \frac{d^2 f}{2!} + \frac{d^3 f}{3!} + \cdots + \frac{d^n f}{n!} + \cdots \tag{62}$$

Truncating this formula we obtain (for small Δx) approximate formulas (more and still more accurate): $\Delta f \approx df$ [accurate to a term of the order of $(\Delta x)^2$], $\Delta f \approx df + \frac{1}{2} d^2 f$ [accurate to a term of the order of $(\Delta x)^3$] and so on.

§ 6. Intervals of Monotonicity. Extremum

17. Sign of Derivative. Let us consider a function $y = f(x)$. We assume throughout this section that the function and its derivative have no discontinuities. A sketch of the graph of the function is shown in Fig. 121. Since $y' = \tan \alpha$ (see Sec. 3) the function increases on every interval where its derivative is positive and decreases on every interval where the derivative is negative. In other

words, if the rate of a variable is positive the variable increases and if the rate is negative the variable decreases.

Since the derivative should pass through the zero value in a continuous transition from its positive values to its negative values there must be $y' = 0$ at those points where an interval of increase borders on an interval of decrease (of course, the same takes place when there is a passage from negative values to positive values of the derivative). A point x at which $f'(x) = 0$ is called a **critical point** of the function $y = f(x)$; the instantaneous rate of change
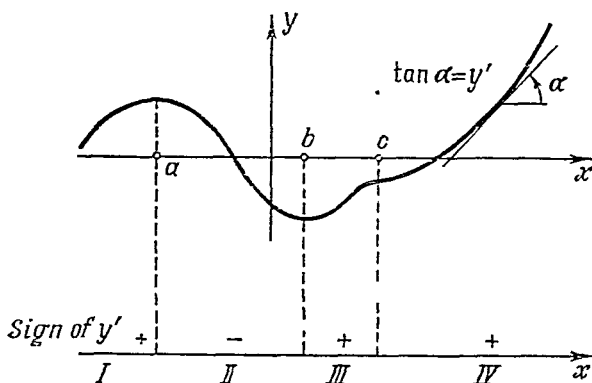
Fig. 121

of the function is equal to zero at such a point, that is critical points serve as if they were "points of instantaneous state of rest". There are three critical points in Fig. 121: a, b and c. The corresponding values of the function are called its **critical** (or **stationary**) **values**.

From what has been said it follows that to determine the intervals of monotonicity of $f(x)$ it is necessary to indicate all the critical points of the function on the x-axis and then to investigate the sign of f' on each interval lying between neighbouring critical points. The intervals where $f' > 0$ will be the intervals of increase of f and the intervals on which $f' < 0$ will be the intervals of decrease of f. In case the sign of f' is the same in two neighbouring intervals these intervals form an entire interval of monotonicity; thus, intervals *III* and *IV* in Fig. 121 constitute a whole interval of increase of the function $f(x)$.

It is also obvious that *the function f is constant on an interval if and only if $f'(x) \equiv 0$ on the interval*. Indeed, the function can neither increase nor decrease on its interval of constancy.

18. **Points of Extremum.** *If the value $f(x_0)$ at some point $x = x_0$ is greater than all the "neighbouring" values of the function* (i.e. greater than the values of $f(x)$ taken for x lying sufficiently close to x_0)

then the point x_0 is called the **point of maximum of the function** $f(x_0)$ (or we say that $f(x_0)$ *has a maximum at the point* $x = x_0$), *and* $f(x_0)$ *is called its* **maximal value. The point of minimum** and the **minimal** value of a function are defined in a similar way. Thus, the function shown in Fig. 121 has a point of maximum at $x = a$ and a point of minimum at $x = b$. In other cases a function can have any other number of points of maximum or minimum and if a function is continuous its maxima and minima must alternate. For example, the function in Fig. 122 has three points of maximum and two points
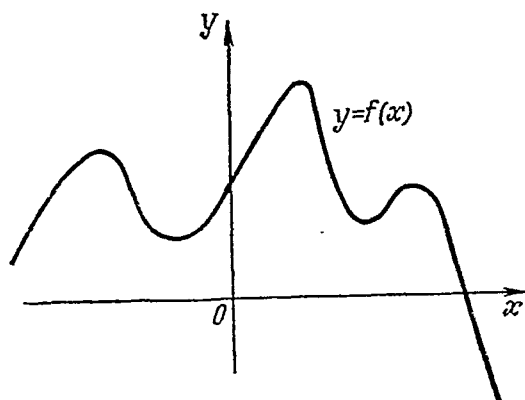
Fig. 122

of minimum; there is an infinitude of maxima and minima in Fig. 46 whereas there are no such points at all in Fig. 44.

A maximum or a minimum of a function is called an **extremum** which means "an extreme value" of the function. From Sec. 17 it follows that points of extremum are such points at which the derivative changes its sign in moving through them from left to right. More definitely, if the sign of $f'(x)$ changes from $+$ to $-$ while x moves through a point $x = a$ in the positive direction then the function f has a maximum at $x = a$ because the function increases on the left of $x = a$ and decreases on the right of $x = a$ (see Fig. 121). Similarly, the derivative changes its sign from $-$ to $+$ in moving through a point of minimum.

It follows that under the conditions specified in Sec. 17 *all the points of extremum of a function are its critical points*. This necessary condition for an extremum was, in fact, formulated by P. Fermat. As it is seen in Fig. 121 this condition is not sufficient, i.e. a critical point may not be a point of extremum.

There are various sufficient conditions for the existence of an extremum but they are used more seldom than the necessary condition because in many concrete problems one often knows beforehand that the extremum must exist. Its approximate location is

also usually known and it is only its exact value that remains unknown. If the necessary condition indicates only one possible point of extremum in the above circumstances then the extremum is sure to be there. In case there are several extrema it is possible to find them and determine the intervals of monotonicity (see Sec. 17) simultaneously.

Since the values of a function change very slowly near a critical point Fermat's condition implies that if a point of extremum is determined with an error then the error of evaluating the corresponding extreme value is of higher order of smallness. Therefore it is usually convenient (if it is possible) to reduce a problem of determining some quantity to the problem of evaluating an extreme or even a stationary (critical) value of a certain function. Then even a rough determination of a point of extremum yields a good ultimate result.

Conditions for an extremum can also be established on the basis of Taylor's formula (see Sec. 15). Let us investigate a point $x = a$ for a function $f(x)$. It is seen from formula (49) that there is no extremum at $x = a$ if $f'(a) \neq 0$ since a change of the sign of h results in the change of the sign of $f'(a) h$ and therefore the sign of the difference $f(a + h) - f(a)$ also changes [because the terms of the order of h^2 are negligibly small compared to $f'(a) h$ for small $|h|$].

If $f'(a) = 0$ and $f''(a) \neq 0$ then, by formula (50), we conclude in like manner that there will be an extremum at $x = a$. Namely, there will be a minimum at $x = a$ provided $f''(a) > 0$ [since in this case $f(a + h) > f(a)$ for small $|h|$] and a maximum provided $f''(a) < 0$. In case $f'(a) = 0$, $f''(a) = 0$ and $f'''(a) \neq 0$ formula (51) implies that there will be no extremum at $x = a$. If $f'(a) = 0$, $f''(a) = 0$, $f'''(a) = 0$ and $f^{IV}(a) \neq 0$ then an extremum exists again and so on.

19. The Greatest and the Least Values of a Function. Let a function $y = f(x)$ and its derivative be continuous on a closed interval $a \leqslant x \leqslant b$. Suppose it is necessary to find the greatest and the least values of the function. In Sec. 18 we considered extrema which were attained at **interior** points of the intervals. But here we must also take into account end-point extrema. For instance, the function in Fig. 123 has a minimum at the end-point $x = a$ and a maximum at the end-point $x = b$ but it also has two other extrema in the interior of the interval. Of course, the derivative of a function is not necessarily equal to zero at an end-point even if there is an extremum there.

Further, it should be noted that in Sec. 18 we investigated relative extrema whereas now we are interested in the absolute maximum and minimum. Therefore in order to find the greatest value of the function on the interval $a \leqslant x \leqslant b$ it is necessary to determine all its maxima at interior points as well as its end-point maxima on the

interval and then to compare the corresponding maximal values with each other: the greatest of these maxima is just the greatest value of the function. The same procedure yields the least value of a function on a closed interval. To facilitate these calculations we can simply compare all the critical and end-point values of the function with each other: the greatest of the values will be the absolute maximum and the least one will be the absolute minimum.

A continuous function $f(x)$ may have a derivative f' with discontinuities. In such a case the transition from the increase of f to its decrease may take place not only at those points where $f' = 0$
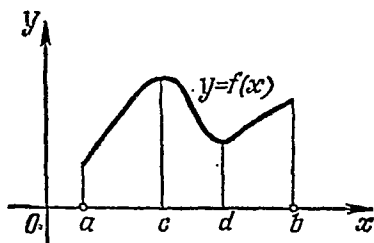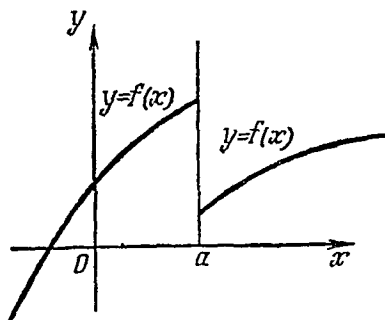
Fig. 123   Fig. 124

but also at some points of discontinuity of the function $f'(x)$. In order to determine the intervals of monotonicity of f we should investigate the sign of f' and this can be carried out in the way the sign of f was determined in Sec. III.15. If the derivative has a discontinuity at a point $x = a$ and changes its sign in the process of moving through the point $x = a$ the function $f(x)$ has a **cuspidal extremum** at $x = a$ (see the cuspidal minimum in Fig. 35 and the cuspidal maxima in Fig. 106). A function f no longer changes slowly near its cuspidal extremum whereas it does change slowly near extrema attained at its critical points (see Sec. 18) where $f' = 0$.

Thus, the comprehensive statement of the necessary condition for the existence of an extremum is the following: *at a point of extremum the derivative vanishes or has a discontinuity.*

The conditions for an extremum based on Taylor's formula no longer hold for the points of discontinuity of a derivative. Only the condition based on the change of the sign of a derivative remains true in such a case.

If a function $f(x)$ itself has discontinuities the points of discontinuity may happen to be the end-points of some intervals of monotonicity of the function even in those cases when the derivative has the same sign on both sides from the point of discontinuity. For instance, as it is seen in Fig. 124, $y' > 0$ everywhere for $x \neq a$ and

the necessary condition for an extremum we obtain

$$\frac{dV}{dx} = 2(a - 2x)(-2)x + (a - 2x)^2 \cdot 1 = (a - 2x)(a - 6x) = 0$$

which implies $x_1 = \frac{a}{2}$ and $x_2 = \frac{a}{6}$. The conditions of the problem indicate that only $x = \frac{a}{6}$ will do. that is this value yields the sought-for maximal volume which is equal to

$$V_{max} = \left(a - 2 \cdot \frac{a}{6}\right)^2 \frac{a}{6} = \frac{2}{27}a^3$$

Example 3. Let us consider the problem of *refraction of the light* passing through the interface between two *homogeneous* (i.e. with the same properties at all points) and *isotropic* (i.e. with the same properties along all directions) media. Let us suppose first that
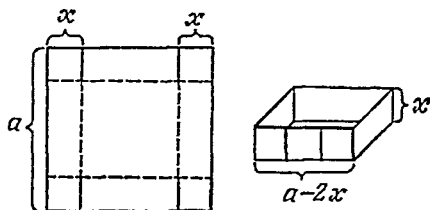
Fig. 127
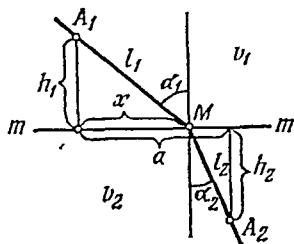
Fig. 128

v_1 is the speed of light in the first medium; v_2 is the speed of light in the second medium; mm is the interface between the media

the interface is plane. Draw a plane through the ray of light (see Fig. 128) and take points A_1 and A_2 on the ray. Now we can use so-called *Fermat's principle in optics* which states that the ray of light propagating from A_1 to A_2 follows a trajectory such that takes the minimal interval of time for passing from A_1 to A_2 in comparison with all the possible trajectories connecting A_1 and A_2. According to the principle the point M (A_1 and A_2 are regarded as fixed) must be located in such a position that the time interval

$$t = t_1 + t_2 = \frac{l_1}{v_1} + \frac{l_2}{v_2} = \frac{\sqrt{h_1^2 + x^2}}{v_1} - \frac{\sqrt{h_2^2 + (a - x)^2}}{v_2}$$

should be as short as possible. Applying the necessary condition for the existence of an extremum we obtain

$$\frac{dt}{dx} = \frac{x}{v_1\sqrt{h_1^2 + x^2}} - \frac{a - x}{v_2\sqrt{h_2^2 - (a - x)^2}} = 0$$

from which it follows that

$$\frac{x}{l_1 v_1} = \frac{a - x}{l_2 v_2}, \quad \frac{\sin \alpha_1}{\sin \alpha_2} = \frac{x}{l_1} : \frac{a - x}{l_2} = \frac{v_1}{v_2}$$

Thus, we have deduced the well-known **law of refraction**: the sine of the angle of incidence bears the constant ratio to the sine of the angle of refraction equal to the ratio of the speeds of light in both media. If now the interface is not plane the law of refraction will remain the same since the phenomenon of refraction depends only on the properties of the media in an infinitesimal vicinity of the point of refraction and the interface can be regarded as being plane in such a vicinity.

Thus we see that we have managed to deduce a physical law on the basis of solving an extremum problem according to a general physical principle formulated in terms of an extremum. Such a principle states that a certain physical quantity must have an extremal value in real circumstances.

A more sophisticated investigation shows that in Fermat's principle (and in some other principles of this kind) the essential condition is not that the time taken by the light to pass a distance must be minimal or even extremal but that the time must assume a stationary value. In the latter form Fermat's principle can be deduced from the wave theory of light.

§ 7. Constructing Graphs of Functions

Differential calculus gives us a general method of investigation of individual peculiarities of the graph of a given function which enables us to construct the graph more accurately and considerably faster than by the primitive method of plotting separate points of the graph as it was done in Sec. I.14. The determination of intervals of monotonicity of a function described in § 6 is an important example of the method of constructing graphs. Besides, there are some other techniques for investigating graphs which are also of use. They will be discussed here.

20. Intervals of Convexity of a Graph and Points of Inflection. Let a function $y = f(x)$ have a graph of the shape shown in Fig. 129. We see that on the left of the point A and on the right of the point B the graph is convex upwards and it is convex downwards between A and B (see Sec. I.24). The points A and B at which the convexity changes its character (i.e. the upward convexity transits to the downward one or vice versa) are the points of inflection; the graph intersects the tangents at these points though it forms the zero angles with them.

In order to find the intervals of upward and downward convexity note that the tangent to a graph turns clockwise as x increases on every interval where the graph is convex upwards (for example, for $x < a$ in Fig. 129) and therefore the slope of the tangent decreases. This slope being equal to y', the graph is convex upwards or downwards on the intervals of the x-axis where y', respectively, decreases

or increases. These intervals can be found by investigating the sign of y'' in the same way as the intervals of decrease and of increase of y were investigated by determining the sign of y' in Sec. 17.
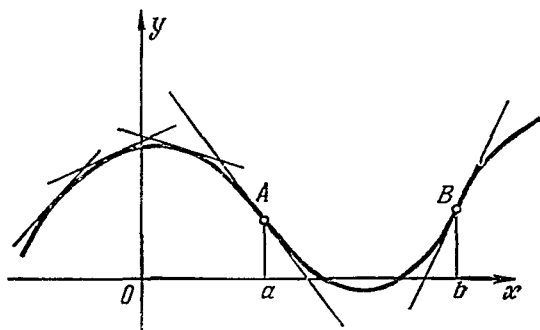
Fig. 129

Therefore, *the graph is convex upwards or downwards on those intervals of the x-axis where $y'' < 0$ or $y'' > 0$, respectively. The points of inflection correspond to values of x such that in moving through x the second derivative y'' changes its sign and y'' is equal to zero at the points of inflection.* It is supposed here that y, y' and y'' have no discontinuities. If there are such discontinuities some of the points of discontinuity may be the end-points of certain intervals of convexity (upward or downward) and therefore all the discontinuities must be indicated on the x-axis while constructing the graph.
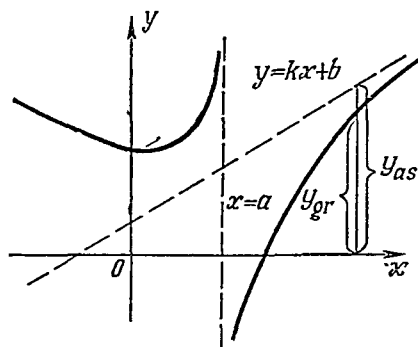
Fig. 130

21. Asymptotes of a Graph. A graph of $y = f(x)$ may have vertical (i.e. parallel to the y-axis) and inclined (i.e. not parallel to the y-axis) asymptotes (see Fig. 130). There may be any number of vertical asymptotes, even an infinite number (see, for example, the tangent curve in Fig. 47), and they are determined in the following way: if $|y| \to \infty$ as $x \to a$ (a is finite) then the straight line $x = a$ is a vertical asymptote.

There cannot be more than two inclined asymptotes corresponding to $x \to \infty$ and $x \to -\infty$ and they are found as follows: let a straight line $y = kx + b$ be an asymptote of the graph of $y = f(x)$ as

$x \to \infty$. Then (see Fig. 130) the difference $\delta = y_{as} - y_{gr}$ is equal to

$$\delta = (kx + b) - f(x) \qquad (63)$$

and tends to zero as $x \to \infty$. Whence,

$$\frac{f(x)}{x} = k + \frac{b}{x} - \frac{\delta}{x} \to k$$

that is

$$k = \lim_{x \to \infty} \frac{f(x)}{x}.$$

Besides, by (63),

$$f(x) - kx = b - \delta \to b$$

that is

$$b = \lim_{x \to \infty} [f(x) - kx]$$

Each of these limits must exist and be finite, otherwise there will be no asymptote as $x \to \infty$. If these finite limits do exist then the asymptote also exists since it is seen from the last equality that the value $[f(x) - kx] - b$ tends to zero as $x \to \infty$, i.e.

$$\lim_{x \to \infty} [f(x) - (kx + b)] = 0$$

22. General Scheme for Investigating a Function and Constructing Its Graph. This scheme for a function $y = f(x)$ consists of the following rules.

(1) We find the domain of definition of the function, its points of discontinuity and zeros and then determine its intervals of positivity and negativity. After that we investigate the behaviour of the function as its argument approaches the points of discontinuity and the end-points of intervals of definition (including the behaviour of the function at infinity). Further, we determine the asymptotes of the graph. We also find out whether the function is even, odd or periodic and so on.

(2) We determine the points of discontinuity and zeros of the derivative and then find the intervals of increase and decrease of the function, its points of extremum and the extremal values. Further, we investigate the behaviour of the derivative in approaching its points of discontinuity, the points of discontinuity of the function itself (in case the function has finite jumps at these points) and the end-points of the intervals on which the function is defined (if these end-points are finite and the function has finite values at them).

(3) We next determine the points of discontinuity of the second derivative and its zeros and then find the intervals on which the function is convex upwards or downwards and also the points of inflection. It is also useful to determine the direction of the tangent at the points of inflection.

All the points thus found should be plotted in the coordinate plane and then the graph itself is constructed. The shape of the graph must reflect all the individual peculiarities of the behaviour of the function. If the elements of the graph indicated above do not describe the behaviour of the graph clear enough it is desirable to plot several additional points by calculating the values of y for some specifically chosen values of x. It is also expedient to determine the direction of the tangent at those points after computing the corresponding values of y'.

We shall represent, as an example, the investigation of the behaviour of the graph of the function $y = \sqrt[3]{x^3 - 2x^2}$. In this case the domain of the function is the whole x-axis $-\infty < x < \infty$; there are no points of discontinuity. Putting $y = 0$ we see that the function has two zeros: $x_1 = 0$ and $x_2 = 2$. Therefore there are three intervals of retention of the sign: $-\infty < x < 0$, $0 < x < 2$ and $2 < x < \infty$. Substituting arbitrary values of the argument taken from these intervals we see that the function is negative on the first and the second intervals and positive on the third one. There are no vertical asymptotes. Determining inclined asymptotes in accordance with Sec. 21 we find (the reader should check it up!) that one and the same straight line $y = x - \frac{2}{3}$ serves as an inclined asymptote both for $x \to \infty$ and $x \to -\infty$. After computing the derivative

$$y' = \frac{3x^2 - 4x}{3\sqrt[3]{(x^3 - 2x^2)^2}} = \frac{3x - 4}{3\sqrt[3]{(x - 2)^2 x}}$$

we see that it has discontinuities (approaches infinity) as $x \to 0$, $x \to 2$ and vanishes at $x = \frac{4}{3}$. Thus we have four intervals of monotonicity: $-\infty < x < 0$, $0 < x < \frac{4}{3}$, $\frac{4}{3} < x < 2$ and $2 < x < \infty$. Substituting arbitrary values from these intervals into y' we find that only the second one is an interval of decrease and all the other intervals are intervals of increase. Therefore the third and the fourth ones form an entire interval of increase. Hence, changes of the character of monotonicity occur at $x = 0$ (where there is a maximum with the maximal value $y = 0$) and at $x = \frac{4}{3}$ (where there is a minimum with the minimal value $y = -2\sqrt[3]{\frac{4}{3}} = -1.058$).

Computing the second derivative we obtain, after some transformations, the expression

$$y'' = -\frac{8}{9\sqrt[3]{(x - 2)^5 x^4}}$$

(verify this expression!). The only discontinuities of the second derivative are the points $x = 0$ and $x = 2$ where the discontinuities

of the first derivative are placed, and there are no zeros of the second derivative at all. Thus, there are three intervals inside which "the character of convexity" is invariable: $-\infty < x < 0$, $0 < x < 2$ and $2 < x < \infty$. Now we substitute arbitrary values taken from these intervals into the second derivative, and then the sign of the derivative shows that the function is convex downwards on the first and on the second intervals and is convex upwards on the third one. Let us, in addition, calculate for $x = -1$ the values

$$y = -\sqrt[3]{3} = -1.44 \quad \text{and} \quad y' = \frac{7}{3\sqrt[3]{9}} = 1.12$$

for $x = 1$,

$$y = -1 \quad \text{and} \quad y' = -\frac{1}{3} = -0.33$$

and for $x = 3$,

$$y = \sqrt[3]{9} = 2.08 \quad \text{and} \quad y' = \frac{5}{3\sqrt[3]{3}} = 1.16$$

The constructed graph is shown in Fig. 131 where those points which were calculated are depicted as circles (we leave it to the
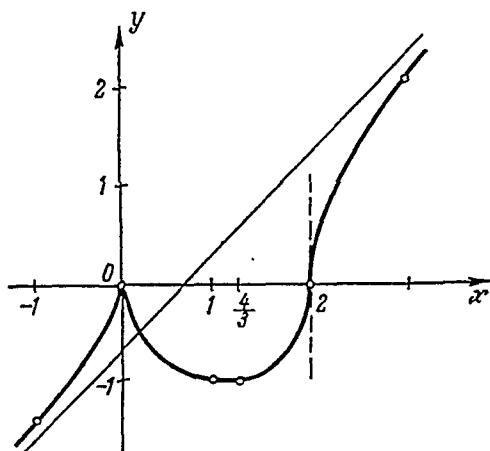
Fig. 131

reader to check that all the individual peculiarities of the graph are indeed represented here).

The disposition of the graph relative to its asymptote can be easily found on the basis of its **asymptotic expansion**, that is an expansion which holds for sufficiently large values of $|x|$. This

expansion, in its turn, follows from formula (60):

$$y = \sqrt[3]{x^3 - 2x^2} = \sqrt[3]{x^3 \left(1 - \frac{2}{x}\right)} = x \left(1 - \frac{2}{x}\right)^{\frac{1}{3}} =$$

$$= x \left[1 + \frac{\frac{1}{3}}{1!} \left(-\frac{2}{x}\right) + \frac{\frac{1}{3}\left(\frac{1}{3} - 1\right)}{2!} \left(-\frac{2}{x}\right)^2 + \ldots\right] =$$

$$= x - \frac{2}{3} - \frac{4}{9x} + \textit{infinitesimal terms of higher order}$$
$$\textit{relative to } \frac{1}{x} \textit{ as } |x| \to \infty \tag{64}$$

Hence, we have $y < x - \frac{2}{3} = y_{as}$ for large $x > 0$ and $y > y_{as}$ for large $x < 0$ as $|x| \to \infty$ (y_{as} denotes here the ordinate of the asymptote). Besides, from equality (64) it straightway follows that

$$y - \left(x - \frac{2}{3}\right) \xrightarrow[|x| \to \infty]{} 0 \tag{65}$$

Hence, if we had not known the equation of the asymptote $y = x - \frac{2}{3}$ before we could deduce it from (65). Therefore, we have established one more method of finding an inclined asymptote.

Approximating Roots of Equations. Interpolation

§ 1. Approximating Roots of Equations

1. Introduction. We shall discuss here some methods of calculating a numerical solution of an equation of the form

$$f(x) = 0 \qquad (1)$$

where f is a given function. Such an equation may be **algebraic** in case the function f is algebraic or **transcendental** if otherwise. We shall call both types of equation (1) finite to distinguish them, for example, from differential equations etc. Here we shall represent only some of the most important methods of solving equations of form (1); for other methods the reader is referred to special courses on calculus of approximations.

The process of numerical solution of equation (1) usually begins with finding a rough, approximate value of a root which is called the **zeroth approximation** (the **initial approximation**). If a certain physical problem is being considered such an initial approximation may be known from the real physical conditions of the problem. We can also begin with constructing an approximate sketch of the graph of the function f. If doing this we find that the function is continuous over a closed interval between a and b and assumes values of opposite signs at the end-points a and b then we are sure, by the properties of continuous functions (see Sec. III.14), that there exists at least one zero of f on the interval, that is equation (1) has at least one root there. Besides, there must be only one root of f on the interval provided the function f is monotonic between a and b. In the last case the root is separated from other ones. If we denote the unknown root by α then α is sure to satisfy the inequality $a < \alpha < b$. Different methods are used for further specification of the value α (see Sec. 2).

It is sometimes more convenient to rewrite equation (1) in the form $\varphi(x) = \psi(x)$ and then to find the intersection point of the graphs of $y = \varphi(x)$ and $y = \psi(x)$. Doing this one tries to break the left-hand side of equation (1) into two summands in such a way that

this should yield some well-known or, at any rate, simpler graphs. An appropriate substitution for the variable x may also be sometimes of use.

For instance, let us take the equation

$$\tan ax^2 - bx^2 = 0 \tag{2}$$

where a and b are given numbers. The change $ax^2 = s$ reduces (1) to the equation

$$\tan s = ks \quad \left(k = \frac{b}{a}\right) \tag{3}$$

The graphs of the left-hand and right-hand sides of equation (3) are shown in Fig. 132. It is clear that we are interested in the values $s \geqslant 0$ only. We see that equation (3) has an infinitude of roots
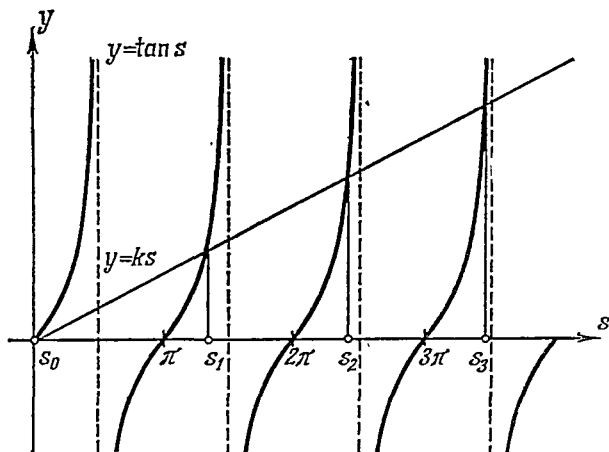
Fig. 132

$s_0 = 0 < s_1 < s_2 < \ldots$, and therefore equation (2) also has infinitely many roots. The dependence of the roots of (3) on k which can be easily seen in Fig. 132 defines the dependence of the roots of the original equation on the parameters a and b. We see that for $k > 1$ there appears a new root lying in the interval $0 < s < \frac{\pi}{2}$ (why is it so?).

We can easily derive an *asymptotic expression* for the solution of equation (3) valid for large n. For definiteness, let $k < 1$. Then Fig. 132 implies the desired expression $s_n = n\pi + \frac{\pi}{2} - \alpha_n$ $(\alpha_n > 0)$ where $\alpha_n \to 0$ as $n \to \infty$; this can be rewritten (see the

notation of Sec. III.11) as $s_n = n\pi + \frac{\pi}{2} + o\,(1)$. If it is necessary to specify this expansion we can substitute it into (3) which results in

$$\tan\left(n\pi + \frac{\pi}{2} - \alpha_n\right) = k\left(n\pi + \frac{\pi}{2} - \alpha_n\right)$$

and after simple transformations we obtain

$$\cos\alpha_n = k\left(n\pi + \frac{\pi}{2} - \alpha_n\right)\sin\alpha_n \tag{4}$$

From this we deduce

$$\alpha_n \sim \sin\alpha_n = \frac{\cos\alpha_n}{k\left(n\pi + \frac{\pi}{2} - \alpha_n\right)} \sim \frac{1}{k\pi n}, \quad \text{i.e.} \quad \alpha_n = \frac{1'}{k\pi n} + o\left(\frac{1}{n}\right)$$

If further specification is desired then we may, for example, denote $\frac{1}{n} = t \to 0$, $\alpha_n = \alpha\,(t) \xrightarrow[t \to 0]{} 0$, which yields

$$t\cos\alpha = k\left[\pi + \left(\frac{\pi}{2} - \alpha\right)t\right]\sin\alpha \quad (\alpha = \alpha\,(t);\ \alpha\,(0) = 0)$$

Now it is possible to put down several terms of the expansion of $\alpha\,(t)$ into Maclaurin's series [of form (IV.54) but in powers of t]. The calculations which we leave to the reader give

$$\alpha = \frac{1}{k\pi}\,t - \frac{1}{2k\pi}\,t^2 + \frac{1}{k\pi}\left(\frac{1}{4} + \frac{1}{k\pi^2} - \frac{1}{3k^2\pi^2}\right)t^3 + \ldots =$$

$$= \frac{1}{k\pi n} - \frac{1}{2k\pi n^2} + \ldots$$

Now on the basis of formula (IV.60) we obtain an asymptotic expression for the positive roots of equation (2) for large n:

$$x_n = \sqrt{\frac{s_n}{a}} = a^{-\frac{1}{2}}\sqrt{n\pi + \frac{\pi}{2} - \frac{1}{k\pi n} + \frac{1}{2k\pi n^2} - \ldots} =$$

$$= \left(\frac{n\pi}{a}\right)^{\frac{1}{2}}\left(1 + \frac{1}{2n} - \frac{1}{k\pi^2 n^2} + \ldots\right)^{\frac{1}{2}} =$$

$$= \left(\frac{n\pi}{a}\right)^{\frac{1}{2}}\left[1 + \frac{1}{4n} - \left(\frac{1}{2k\pi^2} + \frac{1}{32}\right)\frac{1}{n^2} + \ldots\right]$$

2. Cut-and-Try Method. Method of Chords. Method of Tangents.

We begin with the **cut-and-try method**. Its scheme is the following. Let, for definiteness, $f\,(a) < 0$ and $f\,(b) > 0$. We first take an arbitrary value c between a and b and compute $f\,(c)$. It should be noted that it is the sign of $f\,(c)$ that is important here but not the value $f\,(c)$ itself. Now let us suppose that we obtain $f\,(c) > 0$. This means that we have "a shot over the target", "a plus round", and therefore $a < \alpha < c$. Further, we take some value d between a and c and

compute $f(d)$; if $f(d) < 0$ then we have "a minus round", i.e. $d <$ $< \alpha < c$ and so on. The values c, d, ... may be taken more or less arbitrarily. It is better, of course, to choose them in such a way that the calculations should be simpler. At the same time it should be noted that if, for example, $|f(a)|$ is much smaller than $f(b)$, it is quite likely that α is closer to a than to b and therefore it is better to take c closer to a and so on.

The **method of chords** prescribes to take for the point c not an arbitrary point but the point of intersection of the chord drawn through the points $M[a, f(a)]$ and $N[b, f(b)]$ with the x-axis (see Fig. 133). In other words, we act as if we approximately replaced the arc of the graph by a line segment. This means that we are carrying out the **linear interpolation** which looks sufficiently justified provided the interval (a, b) is not too large. In order to find the point c let us write the equation of the chord MN [see equation (II.22)]:

$$\frac{y - f(a)}{f(b) - f(a)} = \frac{x - a}{b - a}$$

Now putting $y = 0$ we obtain the corresponding value $x = c$:

$$c = a - \frac{f(a)(b-a)}{f(b) - f(a)} = b - \frac{f(b)(b-a)}{f(b) - f(a)} \qquad (5)$$

The procedure can be repeated several times if it is necessary (see Fig. 133).

In the **method of tangents** (also called **Newton's method**) we choose the point of intersection of the tangent line drawn to the graph through one of the end-points of the considered arc with the x-axis as the point c. The equation of the tangent shown in Fig. 134 has the form [see equation (IV.5)]

$$y - f(b) = f'(b)(x - b)$$

From this, putting $y = 0$, we derive

$$c = b - \frac{f(b)}{f'(b)} \qquad (6)$$

The procedure described here may also be repeated several times (see Fig. 134).

Newton's method may be interpreted irrespective of its geometric meaning. Let us denote the zeroth approximation as x_0 and expand the left-hand side of (1) in powers of $x - x_0$ on the basis of Taylor's ormula (IV.53); this yields the equation

$$f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots = 0$$

If now we carry out the linearization, that is if we drop all the terms that are infinitesimals of orders higher than the first, we shall get

the linearized equation (1):

$$f(x_0) + f'(x_0)(x - x_0) = 0$$

The solution of this equation is

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

It can be taken as the first approximation of the root of equation (1). Thus we arrive at the same formula (6). The second approximation can be obtained from the first approximation by using the formula

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \tag{7}$$

and so forth. Newton's method always leads to the aim provided the zeroth approximation does not lie too far from the desired root.

The following modification of Newton's method is sometimes used: the denominators in formula (7) and in the formulas for the
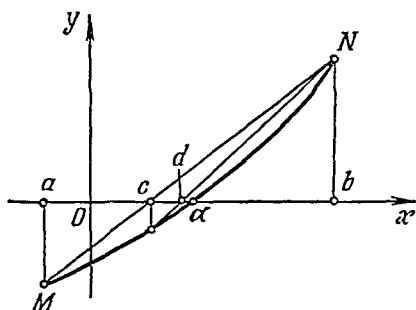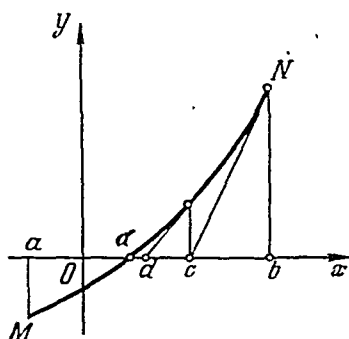
Fig. 133   Fig. 134

further approximations are replaced by $f'(x_0)$. This means, geometrically, that all the inclined lines in Fig. 134 are drawn so that they are parallel to the tangent at the original point N. The convergence of the modified method is a little worse compared with the original scheme but the calculation of each approximation is, naturally, simplified.

The combined method is based on the following consideration. If the segment of the graph in question has no points of inflection and is not broken the method of chords and the method of tangents give the points lying on different sides of the desired root. If, for example, a graph is situated in the way shown in Fig. 135 then, beginning with the interval (a, b), we can determine the point a_1 by the method of chords and the point b_1 by the method of tangents. This will result in a new interval (a_1, b_1) containing the desired

root α. Repeating the analogous procedure for the interval (a_1, b_1) we again obtain a new interval (a_2, b_2) containing the desired root etc. Thus, we obtain in succession the two-sided approximation The approximation process is stopped when the required degree accuracy is attained.

Let us take, for example, the equation

$$x^3 + x^2 - 3 = 0 \qquad ($$

We shall consider the coefficients of equation (8) to be quite exact The investigation of the derivatives indicates that the left-hand
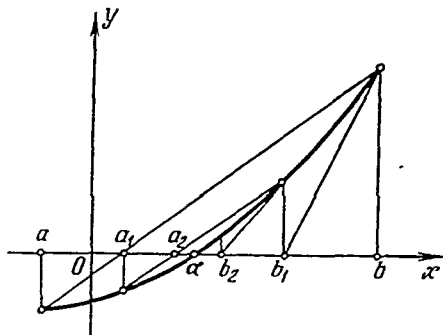
Fig. 135
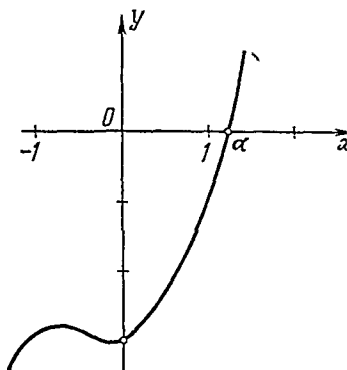
Fig. 136

side of (8) which we denote by $f(x)$ increases from $-\infty$ to $-2\frac{2}{2}$ for x increasing in the interval $-\infty < x < -\frac{2}{3}$, decreases to $-$ for $-\frac{2}{3} < x < 0$ and then again increases to ∞ for $0 < x < \infty$ Besides, $f(x)$ has only one point of inflection at $x = -\frac{1}{3}$ (check up!). Consequently, the equation has a unique real and positive root α. Since $f(0) = -3$, $f(1) = -1$ and $f(2) = 9$ (see Fig. 136) we have $1 < \alpha < 2$. Calculating in accordance with the cut-and-try method we obtain $f(1.1) = -0.459$ and $f(1.2) = 0.168$ i.e. $1.1 < \alpha < 1.2$ (a crude estimation of a root is usually obtained by the cut-and-try method). Now putting $a = 1.1$ and $b = 1.2$ we apply formulas (5) and (6) (i.e. we use the combined method)

$$a_1 = 1.2 - \frac{0.168 \times 0.1}{0.168 + 0.459} = 1.173,$$

$$b_1 = 1.2 - \frac{0.168}{6.72} = 1.175$$

Hence, we can put $\alpha = 1.174$, this value being accurate to 0.001. If the accuracy is insufficient we can proceed to further calculations: $f(1.174) = -0.003628$ (i.e. we have a "minus round", the calculations being accurate to 10^{-6}) and $f(1.175) = 0.002859$. Taking now $a = 1.174$. $b = 1.175$ and calculating by the combined method we obtain the values $a_2 = 1.1745593$ and $b_2 = 1.1745596$ accurate to 10^{-7}. Hence, with an accuracy of 0.000001 we can assume $\alpha = 1.174559$. Note how fast the degree of accuracy increases!

3. **Iterative Method.** The methods described in Sec. 2 belong to the class of iterative methods (or, in other words, to the class of methods of successive approximations). A characteristic feature of all these methods is the successive iteration of one and the same scheme during the calculation process. This uniformity, i.e. the repetition of one and the same process, has many advantages. In particular, it is very convenient when we use digital computers.

The general form of the iterative method applicable to equation (1) is the following: the equation is rewritten in the equivalent form

$$x = \varphi(x) \qquad (9)$$

Then we choose a certain value $x = x_0$ as the zeroth approximation. It is desirable, of course, that x_0 should be as close as possible to the sought-for root if we have some information about it. The subsequent approximations are computed by the formulas $x_1 = \varphi(x_0)$, $x_2 = \varphi(x_1)$, . . ., or, generally,

$$x_{n+1} = \varphi(x_n) \qquad (10)$$

There can be two cases here.

(1) The process may converge, that is the successive approximations x_n tend to a limit $\bar{x}$ as $n \to \infty$. In this case we can pass to the limit in formula (10) which yields $\bar{x} = \varphi(\bar{x})$. Thus, we see that $x = \bar{x}$ is a root of equation (9).

(2) The process may diverge, that is there can be no finite limit for the "approximations" thus constructed. But this fact does not necessarily imply that there is no solution of equation (9) because it might simply occur that the iterative process was constructed in an inappropriate way. By the way, it may happen that even in the case of convergence we obtain some other solution (which may have no physical meaning) quite different from the desired root in whose vicinity x_0 has been chosen.

We shall demonstrate these possibilities by taking an example of a very simple equation which can be easily solved:

$$x = \frac{x}{2} + 1 \qquad (11)$$

The equation has an obvious solution $\bar{x} = 2$. If we put $x_0 = 0$ and calculate with an accuracy of 0.001 we shall obtain $x_1 = 1.000$,

$x_2 = 1.500$, $x_3 = 1.750$, $x_4 = 1.875$, $x_5 = 1.938$, $x_6 = 1.969$, $x_7 = 1.984$, $x_8 = 1.992$, $x_9 = 1.996$, $x_{10} = 1.998$, $x_{11} = 1.999$, $x_{12} = 2.000$ and $x_{13} = 2.000$, that is the process has "practically converged".

If we take the equation

$$x = \frac{x}{10} + 1$$

instead of (11) and assume $x_0 = 0$ then we receive the values $x_1 = 1.000$, $x_2 = 1.100$, $x_3 = 1.110$, $x_4 = 1.111$ and $x_5 = 1.111$ accurate to 0.001; thus, the process practically converged after four iterations.

If we solve equation (11) for x entering into the right-hand side, that is if we rewrite (11) in the equivalent form

$$x = 2x - 2 \qquad (12)$$

and begin with $x_0 = 0$, we shall get the sequence $x_1 = -2$, $x_2 = -6$, $x_3 = -14$ etc., that is the process will not converge. We could have forecast this result if we had observed that formula (10) implied the equality

$$x_{n+1} - x_n = \varphi(x_n) - \varphi(x_{n-1}) \qquad (13)$$

that is $x_2 - x_1 = \varphi(x_1) - \varphi(x_0)$, $x_3 - x_2 = \varphi(x_2) - \varphi(x_1)$ and so forth. In case a function changes more slowly than its argument or, more precisely, if

$$| \varphi(x) - \varphi(\widetilde{x}) | \leqslant k | x - \widetilde{x} | \quad (k = \text{const} < 1) \qquad (14)$$

the distance between the successive approximations rapidly approaches zero and the iterative process converges. The smaller k, the greater the speed of convergence. Inequality (14) must be fulfilled for all x and $\widetilde{x}$ or, at any rate, near the sought-for root $\overline{x}$ of equation (9). We shall show in Sec. 4 that inequality (14) holds provided $| \varphi'(x) | \leqslant k$.

We see that equations (11) and (12) are equivalent but generate different iterative processes. In other cases equation (1) may be rewritten in form (9) in many different ways, each of these ways generating its own iterative process. Some of these processes may happen to converge fast and therefore are more convenient, some of them may converge slowly and, finally, some of them may simply diverge. In particular, it is easy to verify that if we rewrite equation (1) in the form

$$x = x - \frac{f(x)(b - x)}{f(b) - f(x)}$$

and begin with $x_0 = a$ [see formula (5)] this will yield the method of chords. Similarly, if equation (1) is rewritten in the form

$$x = x - \frac{f(x)}{f'(x)} \qquad 15)$$

then we arrive at the method of tangents.

There exists a comprehensive theoretical investigation of the problem of convergence of the iterative method. But in more complicated problems than the above ones it is often easier to compute several approximations. Then judging by the results we can draw the necessary conclusion as to the convergence of the process without giving any theoretical proof. If we see that an approximation differs from the preceding one by a small quantity (for instance, if their difference is less than the required degree of accuracy) we have every reason to stop the iterative process. At any rate, such a situation shows that the approximation satisfies equation (9) with a good accuracy because $|x_n - x_{n+1}| < h$ implies $|x_n - \varphi(x_n)| < h$.

Fig. 137

4. Formula of Finite Increments. Inequality (14) can be verified by means of the so-called formula of finite increments we are going to deduce here. Let us suppose that a function $y = \varphi(x)$ is continuous over an interval $a \leqslant x \leqslant b$. Consider the graph of the function on the interval and draw the chord MN connecting the end-points of the arc (see Fig. 137). Let the derivative of the function also be continuous. Besides, we suppose, for definiteness, that there is a portion of the graph lying above the chord.

Now we draw a straight line ll parallel to MN and lying above the graph. Imagine that we begin to lower the line so that it should remain parallel to MN. Then there must exist a moment when the straight line touches the graph at a point p. Thus, *there is at least one point lying on a smooth arc such that the tangent to the arc at this point should be parallel to the chord connecting the end-points of the arc.* If now we equate the slopes of the chord and of the tangent we receive the formula

$$\frac{\varphi(b) - \varphi(a)}{b - a} = \varphi'(c), \quad \text{i.e.} \quad \varphi(b) - \varphi(a) = \varphi'(c)(b - a) \qquad (16)$$

where c is a point lying between a and b. Formula (16) is called the formula of finite increments (since the distance from a to b may ton be small) or **Lagrange's theorem.** We must note that the value c
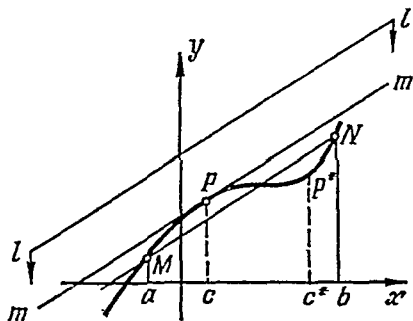
entering into formula (16) is not at all an arbitrary number i
the given function and the given interval (a, b) although there m
be several values that may serve as c. For instance, looking
Fig. 137 we see that we can as well take c^* in place of c in formu
(16) because the tangent at the point p^* is also parallel to the cho
MN. The exact value of c is usually unknown when we apply f
mula (16) but it is often enough to know that c is placed somewhe
between a and b.

For example, suppose that $| \varphi'(x) | \leqslant k$ on an interval. Th
applying formula (16) to two arbitrary points x and $\widetilde{x}$ taken fr
the interval we see that $| \varphi(x) - \varphi(\widetilde{x}) | \leqslant k | x - \widetilde{x} |$ for th
[see formula (14)].

It also follows from formulas (13) and (16) that if the successi
approximations are placed not far from the exact solution $\overline{x}$, a
therefore $\varphi'(x)$ changes very little, the speed of convergence of t
iterative process is approximately that of a geometrical progressi
with the ratio $\varphi'(\overline{x})$. If the differences between successive appro:
mations formed an exact geometrical progression [as in example (1
its first term would be $a = x_1 - x_0$ and the ratio would equal q
$= \frac{x_2 - x_1}{x_1 - x_0}$. Therefore the sum of such a progression, that is t
difference $\overline{x} - x_0$, would be equal to

$$\frac{a}{1-q} = \frac{x_1 - x_0}{1 - \frac{x_2 - x_1}{x_1 - x_0}} = \frac{(x_1 - x_0)^2}{2x_1 - x_0 - x_2}$$

and therefore

$$\overline{x} = x_0 + \frac{(x_1 - x_0)^2}{2x_1 - x_0 - x_2} = \frac{x_1^2 - x_0 x_2}{2x_1 - x_0 - x_2} \qquad (\text{?}$$

In more complicated cases the successive differences of approxim
tions only resemble the terms of a geometrical progression. Th
formula (17) does not yield an exact solution but makes it possi
to omit several approximations and to get an approximate val
of a root which may again initiate a new iterative process.

Newton's iterative process is of special importance. Indeed, 1
derivative of the left-hand side of (15) is equal to

$$1 - \frac{f'f' - ff''}{f'^2} = \frac{ff''}{f'^2}$$

and vanishes at $x = \overline{x}$ since $f(\overline{x}) = 0$. Hence, by the precedi
considerations, Newton's iterative method converges faster th
any geometrical progression with an arbitrary ratio. The rate
this convergence may be easily illustrated by the following typi
example. Let us consider the approximations obtained by Newtor
method for the root $\overline{x} = 0$ of the equation $x + x^2 = 0$. These app

ximations are connected with each other by the relation

$$x_{n+1} = x_n - \frac{x_n + x_n^2}{1 + 2x_n} = \frac{x_n^2}{1 + 2x_n} \approx x_n^2$$

To estimate the rate of convergence let us replace this approximate equality by the exact one. Then we get in succession $x_1 = x_0^2$, $x_2 = = x_1^2 = x_0^4$, $x_3 = x_2^2 = x_0^8$ etc. Generally, $x_n = x_0^{2^n}$. The right-hand side $x_0^{2^n}$ tends to zero as $n \to \infty$, for $|x_0| < 1$, faster than any exponential expression.

5. **Small Parameter Method.** The small parameter method (the perturbation method), as well as the iterative method, is one of the most universal methods in mathematics. Here is the general idea of the method. Let us consider a problem involving some unknown quantities and, in addition, a parameter α. Suppose that it is not too difficult to solve the problem for a certain value $\alpha = \alpha_0$ (this is the so-called *unperturbed solution*). Then the solution for α lying close to α_0 (the so-called *perturbed solution*) may be in many cases obtained by expanding the solution in powers of $\alpha - \alpha_0$, with a certain degree of accuracy, by means of formulas similar to formulas (IV.49), (IV.50), (IV.51) etc. Obviously, the first term of such an expansion does not contain $\alpha - \alpha_0$ and is obtained for $\alpha = \alpha_0$, i.e. it coincides with the unperturbed solution. The subsequent terms yield corrections to the unperturbed solution; these terms are infinitesimals of the first, second etc. orders (relative to $\alpha - \alpha_0$). These terms are usually computed by the method of undetermined coefficients, i.e. the coefficients in $(\alpha - \alpha_0)$, $(\alpha - \alpha_0)^2$ etc. are denoted by letters and the values denoted by the letters are then found on the basis of the conditions of the problem. This method gives a good result only for values of α which are close to α_0. The smaller $|\alpha - \alpha_0|$, the smaller the number of terms that should be computed to attain a desired accuracy. It is often convenient to choose a parameter so that $\alpha_0 = 0$; then the difference $\alpha - \alpha_0 = \alpha$ is considered small and this accounts for the term "the small parameter method". The number of terms that must be taken may be determined by a method similar to the one used in the end of Sec. III.6. It should also be noted that an attempt to use the small parameter method for large $|\alpha - \alpha_0|$ may lead to principal mistakes because the dropped terms may be more significant than the retained ones in this case.

Thus, the small parameter method makes it possible to obtain a solution of a problem that is formulated in terms which are close to the terms of a certain "main" problem provided, of course, this change of the formulation does not yield a principal, qualitative change of the solution. Even determining the first term containing a parameter often enables us to make some useful conclusions concerning the dependence of the solution on the parameter.

Example. Let us solve the equation

$$x^3 - \alpha x^2 + 1 = 0 \tag{18}$$

for small $|\alpha|$ with an accuracy up to the term α^3 inclusively. In order to do this note that the value $\alpha = 0$ yields the equation $x^3 + 1 = 0$ which has an obvious root $x_0 = -1$. Therefore we put down the expression

$x_\alpha = -1 + a\alpha + b\alpha^2 + c\alpha^3 + \textit{infinitesimals of higher order}$

Substituting this expression into (18) and taking terms only up to the order of α^3 we receive (check it!)

$$(-1 + 3a\alpha + 3b\alpha^2 - 3a^2\alpha^2 - 6ab\alpha^3 + 3c\alpha^3 + a^3\alpha^3) -$$
$$- \alpha(1 - 2a\alpha - 2b\alpha^2 + a^2\alpha^2) + 1 +$$
$$+ \textit{infinitesimals of higher order} = 0$$

From this, equating coefficients in the same powers of α, we derive $3a - 1 = 0$, $3b - 3a^2 + 2a = 0$ and $-6ab + 3c + a^3 + 2b - a^2 = 0$. Now we find, in succession, $a = \dfrac{1}{3}$, $b = -\dfrac{1}{9}$ and $c = \dfrac{2}{81}$. Hence, we obtain the expression

$$x_\alpha = -1 + \frac{\alpha}{3} - \frac{\alpha^2}{9} + \frac{2\alpha^3}{81} \tag{19}$$

for the root of equation (18) which is accurate to infinitesimals of higher order relative to α (namely, the error is of the order of α^4).

Just the same result may be obtained by applying directly Taylor's formula (IV.51) in which we change the notation a little:

$$x_\alpha = x_0 + \left(\frac{dx}{d\alpha}\right)_0 \alpha + \frac{1}{2!}\left(\frac{d^2x}{d\alpha^2}\right)_0 \alpha^2 + \frac{1}{3!}\left(\frac{d^3x}{d\alpha^3}\right)_0 \alpha^3 \tag{20}$$

Here the subscript "zero" points out that the value $\alpha = 0$ is substituted into the corresponding terms. Now let us differentiate equality (18) with respect to α in a manner similar to the one used in Sec. IV.11:

$$3x^2 \frac{dx}{d\alpha} - x^2 - 2\alpha x \frac{dx}{d\alpha} = 0,$$

$$6x\left(\frac{dx}{d\alpha}\right)^2 + 3x^2 \frac{d^2x}{d\alpha^2} - 4x \frac{dx}{d\alpha} - 2\alpha\left(\frac{dx}{d\alpha}\right)^2 - 2\alpha x \frac{d^2x}{d\alpha^2} = 0,$$

$$6\left(\frac{dx}{d\alpha}\right)^3 + 18x \frac{dx}{d\alpha} \frac{d^2x}{d\alpha^2} + 3x^2 \frac{d^3x}{d\alpha^3} - 6\left(\frac{dx}{d\alpha}\right)^2 - 6x \frac{d^2x}{d\alpha^2} -$$
$$- 6\alpha \frac{dx}{d\alpha} \frac{d^2x}{d\alpha^2} - 2\alpha x \frac{d^3x}{d\alpha^3} = 0$$

Substituting $\alpha = 0$ and $x = -1$ we derive

$$3\left(\frac{dx}{d\alpha}\right)_0 - 1 = 0, \quad -6\left(\frac{dx}{d\alpha}\right)_0^2 + 3\left(\frac{d^2x}{d\alpha^2}\right)_0 + 4\left(\frac{dx}{d\alpha}\right)_0 = 0,$$

$$6\left(\frac{dx}{d\alpha}\right)_0^3 - 18\left(\frac{dx}{d\alpha}\right)_0\left(\frac{d^2x}{d\alpha^2}\right)_0 + 3\left(\frac{d^3x}{d\alpha^3}\right)_0 - 6\left(\frac{dx}{d\alpha}\right)_0^2 + 6\left(\frac{d^2x}{d\alpha^2}\right)_0 = 0.$$

from which we obtain, in succession,

$$\left(\frac{dx}{d\alpha}\right)_0 = \frac{1}{3}; \quad \left(\frac{d^2x}{d\alpha^2}\right)_0 = -\frac{2}{9} \quad \text{and} \quad \left(\frac{d^3x}{d\alpha^3}\right)_0 = \frac{4}{27}$$

From this it is seen that formula (20) implies expansion (19) which is sufficiently accurate for small $|\alpha|$.

The small parameter method is closely and directly related to the iterative method of Sec. 3. We shall illustrate this connection by taking the same example (18). It is always convenient to have an unperturbed solution equal to zero. To attain this let us make the substitution $x = -1 + y$ which yields

$$-1 + 3y - 3y^2 + y^3 - \alpha + 2\alpha y - \alpha y^2 + 1 = 0$$

i.e.

$$y = \frac{1}{3}\alpha - \frac{2}{3}\alpha y + y^2 + \frac{1}{3}\alpha y^2 - \frac{1}{3}y^3$$

If now we carry out iterations beginning with the value $y_0 = 0$ and dropping the infinitesimal terms of order higher than the third we get the desired result after three iterations. It is also easy to verify that it is permissible to neglect those terms of each approximation which are infinitesimals of an order higher than the number of the approximation.

§ 2. Interpolation

6. Lagrange's Interpolation Formula. As it was shown in Sec. I.22, the process of linear interpolation consists in an approximate replacement of a given function $y = f(x)$ by a linear function $y = ax + b$ coinciding with $f(x)$ at two points. Obviously, the accuracy of such an approximation may be increased by taking a polynomial of degree n of the form

$$P(x) = P_n(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n$$

in place of a linear function.

The polynomial $P_n(x)$ approximating the function $f(x)$ contains $n + 1$ parameters (i.e. its coefficients), and thus $n + 1$ conditions, in general, are needed to determine such a polynomial. Let us take, for the sake of simplicity, a polynomial of the second degree:

$$P(x) = P_2(x) = ax^2 + bx + c$$

(the general case is investigated quite similarly). To choose such a polynomial we must set three conditions. These conditions are often taken in the following form: the polynomial should coincide with the function $f(x)$ at three given points:

$$P(x_1) = f(x_1), \quad P(x_2) = f(x_2), \quad P(x_3) = f(x_3) \qquad (21)$$

These three values are also regarded as known:

It is quite evident that there can be only one such polynomial. Indeed, if another polynomial of the second degree $Q(x)$ satisfied conditions (21) the difference $P(x) - Q(x)$ (which is also a polynomial of the second degree) would be equal to zero at $x = x_1$, $x = x_2$ and $x = x_3$. This implies that the difference is identically equal to zero (why is it so?). Thus, all the coefficients of the difference are equal to zero, and therefore $Q(x) \equiv P(x)$.

Lagrange's idea is to look for a polynomial $P(x)$ in the form

$$P(x) = A(x - x_2)(x - x_3) + B(x - x_1)(x - x_3) +$$
$$+ C(x - x_1)(x - x_2) \qquad (22)$$

where A, B and C are some constants yet unknown. It is clear that this is a polynomial of the second degree. To find the constants A, B and C let us take conditions (21) and notice that substituting $x = x_1$, $x = x_2$ and $x = x_3$ into the right-hand side of formula (22) yields only one nonzero summand while the other two vanish. Hence we obtain

$$f(x_1) = A(x_1 - x_2)(x_1 - x_3), \quad f(x_2) = B(x_2 - x_1)(x_2 - x_3),$$
$$f(x_3) = C(x_3 - x_1)(x_3 - x_2)$$

Now we find A, B and C from the latter relations and substitute them into (22). Thus we have deduced *Lagrange's interpolation formula*

$$f(x) \approx P_2(x) = f(x_1)\frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} +$$
$$+ f(x_2)\frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} + f(x_3)\frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} \qquad (23)$$

For practical applications of the formula it is desirable that none of the differences $x_1 - x_2$, $x_1 - x_3$ and $x_2 - x_3$ should be too small. (Think why this condition is important.)

Conditions (21) may be replaced, for example, by the following three conditions:

$$P(x_1) = f(x_1), \quad P'(x_1) = f'(x_1), \quad P(x_2) = f(x_2)$$

Then the polynomial $P(x)$ may be looked for in the form

$$P(x) = A(x - x_2)(x - 2x_1 + x_2) +$$
$$+ B(x - x_1)(x - x_2) + C(x - x_1)^2$$

instead of (22). (Find the coefficients A, B and C for this case!)

7. **Finite Differences and Their Connection with Derivatives.** Before proceeding to our further investigations let us introduce one of the important notions of modern mathematics, namely, the notion of a finite difference. Let $y = f(x)$. Then, with h given, the expression

$$\Delta_h y = f(x + h) - f(x)$$

is called a **finite difference of the first order of a function** f (the **first difference of** f). The expression

$$\frac{1}{h} \Delta_h y = \frac{f(x+h) - f(x)}{h}$$

is called the **first difference quotient** or the **first divided difference**. It follows from the definition of a derivative (see Sec. IV.2) that for a sufficiently small h we have

$$\frac{1}{h} \Delta_h y \approx y' \tag{24}$$

or, more precisely,

$$y' = \lim_{h \to 0} \frac{1}{h} \Delta_h y$$

Let, for example, $y = x^3$. Then

$$\Delta_h y = (x+h)^3 - x^3 = 3x^2 h + 3xh^2 + h^3$$

$$\frac{1}{h} \Delta_h y = 3x^2 + 3xh + h^2$$

$$\lim_{h \to 0} \left(\frac{1}{h} \Delta_h y \right) = \lim_{h \to 0} (3x^2 + 3xh + h^2) = 3x^2 = y'$$

We also indicate here the following obvious properties of differences:

$$\Delta_h (y_1 + y_2) = \Delta_h y_1 + \Delta_h y_2,$$
$$\Delta_h (Cy) = C \Delta_h y \quad (C = \text{const})$$

We can also take a difference of the first differences, the so-called **second difference**:

$$\Delta_h^2 y = \Delta_h (\Delta_h y) = \Delta_h [f(x+h) - f(x)] =$$
$$= [f(x+2h) - f(x+h)] - [f(x+h) - f(x)] =$$
$$= f(x+2h) - 2f(x+h) + f(x)$$

The **second divided difference** is defined in an analogous way:

$$\frac{1}{h} \Delta_h \left(\frac{1}{h} \Delta_h y \right) = \frac{1}{h^2} \Delta_h (\Delta_h y) = \frac{1}{h^2} \Delta_h^2 y = \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2}$$

Since taking a divided difference with a small step is approximately equivalent to differentiating, the second divided difference for a small step is approximately equal to the second derivative or, more precisely,

$$y'' = \lim_{h \to 0} \frac{1}{h^2} \Delta_h^2 y = \lim_{h \to 0} \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} \tag{25}$$

Thus, in the previous example

$$\Delta_h^2 y = \Delta_h \left(3x^2 h + 3xh^2 + h^3\right) = 3\left(x+h\right)^2 h + 3\left(x+h\right)h^2 + h^3 - 3x^2 h - 3xh^2 - h^3 = 6xh^2 + 6h^3;$$

$$\lim_{h \to 0} \frac{1}{h^2} \Delta_h^2 y = \lim_{h \to 0} \left(6x + 6h\right) = 6x = y''$$

The third difference $\Delta_h^3 y = \Delta_h \left(\Delta_h^2 y\right)$ and the third divided difference $\frac{1}{h^3} \Delta_h^3 y$ (which tends to the third derivative y''' when passing to the limit) etc. are defined similarly.

It is especially convenient for computing the differences when a function is represented by a table with a constant step h. In case a table is given in general form (I.2) we can write $\Delta y_1 = y_2 - y_1$, $\Delta y_2 = y_3 - y_2$ and, generally, $\Delta y_k = y_{k+1} - y_k$. The subscript k in the expression Δy_k now indicates the number of the difference and not the step because the step is regarded as fixed here. Further, $\Delta^2 y_1 = \Delta y_2 - \Delta y_1$, $\Delta^2 y_2 = \Delta y_3 - \Delta y_2$ and so on. For instance, let us take. for $h = 0.1$. the table

x	10.0	10.1	10.2	10.3	10.4	10.5	10.6	10.7
y	1.00000	1.00432	1.00860	1.01284	1.01703	1.02119	1.02531	1.02938
$10^5 \Delta y$	432	428	424	419	416	412	407	
$10^5 \Delta^2 y$	-4	-4	-5	-3	-4	-5		
$10^5 \Delta^3 y$	0	-1	2	-1	-1			

(This fragment is taken from the table of logarithms. The values of the differences are multiplied by 100,000 in order to get rid of decimal zeros.)

The smallness and the approximate constancy of the second differences in the above example indicate the smoothness of the process of change of the function and the absence of random "splashes" in the process. Such a smoothness may be manifested in differences of higher order and it always indicates the "regularity" of the change of a function. In case the step is not small or the values of the argument are close to the points of discontinuity etc. the differences may not be small but, as a rule, a certain kind of regularity in their values can be noticed. At the same time random errors occurring in the table greatly influence differences of higher order and this

usually enables us to find the errors. Suppose that by mistake we wrote 1.01294 instead of 1.01284 in the table. Then the fourth line would look as $-4. +6, -25, +7, -4, -5$ (check it!) and the regularity would obviously be broken. That is why differences of an order higher than the second are rarely used when a table represents results of an experiment which was not carried out with high precision. In such cases one often restricts oneself to the first differences.

The difference $y_{k+1} - y_k$ is sometimes attributed not to the value x_k, as above, but to the value $x_k + \frac{h}{2}$ which is naturally denoted as $x_{k+\frac{1}{2}}$. Then the difference is called **central** and is designated by $\delta_{k+\frac{1}{2}} y = y_{k+1} - y_k$. Dividing the difference by the step we obtain a **divided central difference**. The central differences of the second order $\delta_k^2 y = \delta_{k+\frac{1}{2}} y - \delta_{k-\frac{1}{2}} y$ are formed similarly. They are again attributed to the "integer" values of the argument etc. (of course, it is not the values x_k of the argument x that are integers but their numbers k; the same is with the "half-integer" values $x_{k+\frac{1}{2}}$ etc.).

It is seen in Fig. 138 that the value of the divided central difference which is equal to the slope of the chord BC is closer to the derivative (i.e. to the slope of the tangent at the point A) than the "simple" divided difference (which is equal to the slope of the chord AD). This assertion can be easily verified by Taylor's series

Fig. 138

(IV.52) since the difference

$$\frac{\Delta y}{h} - y' = \frac{y(x+h) - y(x)}{h} - y'(x) = \frac{h}{2} y''(x) + \frac{h^2}{6} y'''(x) + \ldots$$

is of the order of h for small $|h|$ whereas the difference

$$\frac{\delta y}{h} - y' = \frac{y\left(x + \frac{h}{2}\right) - y\left(x - \frac{h}{2}\right)}{h} - y'(x) = \frac{h^2}{24} y'''(x) + \frac{h^4}{1920} y^V(x) + \ldots$$

is of the order of h^2 (check it!). (The estimation of orders of accuracy of other approximate formulas for a small step is carried out similarly.) Thus, it is better to compute the approximate values of a derivative by means of a divided central difference than by formula (24). A more precise method of approximate calculation of derivatives of any order will be given in Sec. 9.

13*
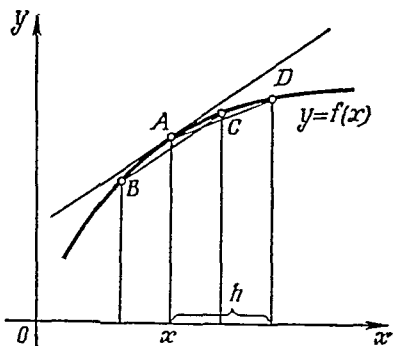
Divided central differences for a small step are close to the corresponding derivatives and resemble them in many respects whereas the differences themselves (not divided) are close to the corresponding differentials. For instance, formula (25) implies

$$\frac{1}{h^2}\,\Delta_h^2 y = y'' + \alpha$$

where $|\,\alpha\,| \ll 1$, i.e. α is an infinitesimal as $h \to 0$. It follows that

$$\Delta_h^2 y = y''h^2 + \alpha h^2 = y''\,(\Delta x)^2 + \alpha h^2 =$$
$$= d^2 y + \alpha h^2 \quad (\,|\,\alpha h^2\,| \ll h^2)$$

Hence, in case $y'' \neq 0$ the value of $\Delta_h^2 y$ differs from that of $d^2 y$ by an infinitesimal of higher order and $\Delta_h^2 y$ and $d^2 y$ are therefore equivalent infinitesimals as $h \to 0$ (see Sec. III.8).

The theory of finite differences was developed simultaneously with other basic branches of mathematical analysis. The first systematic representation of calculus of finite differences was given by Taylor in 1715. Finite differences are nowadays widely used in many theoretical investigations and in practical applications especially in connection with modern electronic computers.

8. Newton's Interpolation Formulas. If the distance h between neighbouring values of x for which a function f is given is constant we can use some formulas that are more convenient than formula (23). For example, suppose we know the values

$$f\,(x_0) = y_0, \quad f\,(x_1) = y_1, \quad f\,(x_2) = y_2, \quad f\,(x_3) = y_3$$

where $x_1 = x_0 + h$, $x_2 = x_0 + 2h$ and $x_3 = x_0 + 3h$. Then the polynomial $P\,(x)$ taking the same values for the appointed values of x will be of the third degree (see Sec. 6). Newton's idea was to seek $P\,(x)$ as a polynomial of the form

$$P\,(x) = A + Bs + Cs\,(s - h) + Ds\,(s - h)\,(s - 2h) \qquad (26)$$

where $s = x - x_0$. According to the condition there must be

$$y_0 = P\,(x_0) = P\,|_{s=0} = A, \quad y_1 = P\,(x_1) = P\,|_{s=h} = A + Bh,$$
$$y_2 = P\,(x_2) = P\,|_{s=2h} = A + B\cdot 2h + C\cdot 2h^2,$$
$$y_3 = P\,(x_3) = P\,|_{s=3h} = A + B\cdot 3h + C\cdot 3\cdot 2h^2 + D\cdot 3\cdot 2h^3$$

Writing differences (see Sec. 7) for the left-hand and right-hand sides we obtain

$$\Delta y_0 = Bh, \quad \Delta y_1 = Bh + C\cdot 2h^2, \quad \Delta y_2 = Bh + C\cdot 2\cdot 2h^2 + D\cdot 3\cdot 2h^3$$

Forming differences a second and a third time we get
$$\Delta^2 y_0 = C\cdot 2h^2, \quad \Delta^2 y_1 = C\cdot 2h^2 + D\cdot 3\cdot 2h^2, \quad \Delta^3 y_0 = D\cdot 3\cdot 2h^2$$

From this we find $A = y_0$, $B = \frac{\Delta y_0}{h}$, $C = \frac{\Delta^2 y_0}{2!h^2}$ and $D = \frac{\Delta^3 y_0}{3!h^3}$. Substituting these values in (26) and taking into account that we could start from any tabular value x_k in place of x_0 we derive Newton's formula

$$f(x) \approx P(x) = y_k + \Delta y_k \frac{s}{h} + \frac{\Delta^2 y_k}{2!} \frac{s}{h} \left(\frac{s}{h} - 1 \right) +$$

$$+ \frac{\Delta^3 y_k}{3!} \frac{s}{h} \left(\frac{s}{h} - 1 \right) \left(\frac{s}{h} - 2 \right) \tag{27}$$

where $s = x - x_k$.

Formulas for interpolation polynomials of other degrees are of a similar form. Increasing this degree we can pass to the limit as it was done in Sec. IV.16 and thus obtain an infinite series of the form

$$f(x) = y_k + \Delta y_k \frac{s}{h} + \frac{\Delta^2 y_k}{2!} \frac{s}{h} \left(\frac{s}{h} - 1 \right) + \frac{\Delta^3 y_k}{3!} \frac{s}{h} \left(\frac{s}{h} - 1 \right) \left(\frac{s}{h} - 2 \right) + \cdots \tag{28}$$

The terms that are not put down in formula (28) contain differences of the fourth, fifth etc. orders and are therefore infinitesimals of the fourth, fifth etc. orders relative to the step h. This formula is, of course, truncated in practical calculations. The number of retained terms must be chosen in such a way that the dropped terms should be negligibly small. It may turn out that it is impossible to attain this in case the step is too large or if we consider points lying near the end-point of the interval. In such circumstances formula (28) is inapplicable.

Newton's formulas (27) and (28) are easy to use when a function f is represented in a tabular way since in such a case its differences are calculated quite simply. They are especially often applied in the beginning of a table (when, for example, we take $k = 0$, i.e. the first tabular value x_0. and $x_0 < x < x_1$). We choose the degree of an interpolation polynomial $P(x)$ considering the values of the differences. For example, if the third differences are very small then the last term in formula (27) is also small and can be dropped. that is we can restrict ourselves to a polynomial of the second degree. In formula (28) it is also possible to put $k = 0$, $x < x_0$ if $|x - x_0|$ is not large; this will result in a *backward extrapolation of the table*.

Newton's another formula

$$f(x) = y_{k+1} - \Delta y_k \frac{t}{h} + \frac{\Delta^2 y_{k-1}}{2!} \frac{t}{h} \left(\frac{t}{h} - 1 \right) -$$

$$- \frac{\Delta^3 y_{k-2}}{3!} \frac{t}{h} \left(\frac{t}{h} - 1 \right) \left(\frac{t}{h} - 2 \right) + \cdots \tag{29}$$

where $t = x_{k+1} - x$ can be deduced like (28). This formula is used, in particular, at the end of a table, for example, if x_{k+1} is the last

tabular value of the argument and $x_h < x < x_{h+1}$. This formula is also used for a *forward extrapolation of a table.*

When interpolation is carried out in the middle of a table it is desirable to have a formula that takes into account tabular values lying on the left and on the right of the value x we are interested in. One of such formulas is *Bessel's formula* which is obtained by taking the half-sum of the right-hand sides of (28) and (29):

$$f(x) = \frac{y_h + y_{h+1}}{2} + \Delta y_h \left(\frac{s}{h} - \frac{1}{2} \right) + \frac{\Delta^2 y_{h-1} + \Delta^2 y_h}{2 \cdot 2!} \frac{s}{h} \left(\frac{s}{h} - 1 \right) +$$
$$+ \frac{\Delta^3 y_{h-1}}{3!} \frac{s}{h} \left(\frac{s}{h} - 1 \right) \left(\frac{s}{h} - \frac{1}{2} \right) + \cdots \tag{30}$$

where $s = x - x_h$. This formula provides a high accuracy; it was named after F. W. Bessel (1784–1846), a German astronomer, but it was in fact established by Newton.

Interpolation formulas are also used for the so-called *problem of inverse interpolation* which consists in finding the values of the argument for given values of a function. Let us begin, for example, with formula (27). Regarding this equality as an exact one we can solve it for the second summand in the right-hand side which, after division by Δy_h, yields

$$\frac{s}{h} = \frac{y - y_h}{\Delta y_h} - \frac{\Delta^2 y_h}{2! \Delta y_h} \frac{s}{h} \left(\frac{s}{h} - 1 \right) - \frac{\Delta^3 y_h}{3! \Delta y_h} \frac{s}{h} \left(\frac{s}{h} - 1 \right) \left(\frac{s}{h} - 2 \right) \quad [(31)$$

If y is given then in order to find s we can apply the iterative method (see Sec. 3). To do this we can choose $\left(\frac{s}{h} \right)_0 = \frac{y - y_h}{\Delta y_h}$ as the zeroth approximation. Substituting this value into the right t-hand side of (31) we obtain $\left(\frac{s}{h} \right)_1$ and so forth. The process converges well for small h.

It should be taken into account that when we interpolate a discontinuous function or a function with a discontinuous derivative the accuracy of an approximation may decrease considerably near the points of discontinuity since the interpolation polynomial itself has no discontinuities. A discontinuity may be imitated by bringing nodes of interpolation very close to each other but it is usually preferable to carry out an interpolation process only on intervals lying between the points of discontinuity.

9. Numerical Differentiation. Numerical differentiation is usually applied when a function whose derivative must be found is defined by a table. This can be carried out as follows: the function is replaced by a polynomial according to the methods of Secs. 6 and 8 and then the derivatives of the polynomial are taken as the approximations to the derivatives of the original function. For example, formula (27)

implies (check it!)

$$f'(x) \approx \frac{\Delta y_k}{h} + \frac{\Delta^2 y_k}{h} \left(\frac{s}{h} - \frac{1}{2} \right) + \frac{\Delta^3 y_k}{2h} \left[\left(\frac{s}{h} \right)^2 - 2 \left(\frac{s}{h} \right) + \frac{2}{3} \right]$$

[Taking formula (28) with a greater number of terms we can get a more accurate result.] In particular, putting $x = x_k$ (i.e. $s = 0$) we derive

$$f'(x_k) \approx \frac{1}{h} \left(\Delta y_k - \frac{\Delta^2 y_k}{2} + \frac{\Delta^3 y_k}{3} \right)$$

A more precise formula may be represented as an infinite series of the form

$$f'(x_k) = \frac{1}{h} \left(\frac{\Delta y_k}{1} - \frac{\Delta^2 y_k}{2} + \frac{\Delta^3 y_k}{3} - \frac{\Delta^4 y_k}{4} + \ldots \right) \tag{32}$$

In like manner we can deduce formulas for the derivatives of the second and subsequent orders. Formulas (29) and (30) (and other interpolation formulas) may also be used for these purposes.

Let us, in particular, put down the following formula implied by formula (30):

$$f'(x_k) = \frac{1}{2h} \left\{ (\Delta y_{k-1} + \Delta y_k) - \frac{1}{6} (\Delta^3 y_{k-2} + \Delta^3 y_{k-1}) + \right.$$

$$\left. + \frac{1}{30} (\Delta^5 y_{k-3} + \Delta^5 y_{k-2}) - \ldots \right\}$$

The subsequent terms of the last formula are, respectively, of the first, third, fifth etc. orders, i.e. the coefficients decrease and the orders of infinitesimals increase here faster than those of the terms of series (32).

If a table represents some results of an experiment then even a small error in a value of a function may lead (after the division by a small step) to finite or even a large error in computing a value of the derivative. The situation is getting still worse when we calculate derivatives of higher order. It is therefore desirable that the step of a table should be considerably greater (e.g. 10 times greater) than the possible error in determining the values of the function. The step should be still greater (e.g. 100 times greater) than the error when we calculate the second derivative and so on. As a result of these difficulties it is usually preferable to use other (empirical) formulas (compare with Sec. I.30) instead of interpolation formulas when we differentiate empirical functions. Since empirical formulas are constructed by taking into account all the peculiarities of experimental data they are considerably more stable with respect to random errors of an experiment.

Interpolation formulas and formulas for numerical differentiation are treated in courses on approximate calculations.

Determinants and Systems of Linear Algebraic Equations

§ 1. Determinants

1. Definition. The concept of a determinant arises when we investigate systems of algebraic equations of the first degree. Let us first take the system of equations

$$\begin{aligned} a_1 x + b_1 y &= d_1 \\ a_2 x + b_2 y &= d_2 \end{aligned} \tag{1}$$

in two unknowns x and y. Solving the system (we leave the calculations to the reader) we get the answer

$$x = \frac{d_1 b_2 - b_1 d_2}{a_1 b_2 - b_1 a_2}, \qquad y = \frac{a_1 d_2 - d_1 a_2}{a_1 b_2 - b_1 a_2} \tag{2}$$

The expression $a_1 b_2 - b_1 a_2$ is called the **determinant of the second order**. It is designated by the symbol $\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}$:

$$a_1 b_2 - b_1 a_2 = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \tag{3}$$

This notation enables us to rewrite formulas (2) in the form

$$x = \frac{\begin{vmatrix} d_1 & b_1 \\ d_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}, \qquad y = \frac{\begin{vmatrix} a_1 & d_1 \\ a_2 & d_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} \tag{4}$$

Let us consider an example of computing a determinant:

$$\begin{vmatrix} 0 & -3 \\ 2 & 1 \end{vmatrix} = 0 \cdot 1 - (-3) \, 2 = 0 + 6 = 6$$

The same process applied to solving the system of equations

$$\begin{aligned} a_1 x + b_1 y + c_1 z &= d_1 \\ a_2 x + b_2 y + c_2 z &= d_2 \\ a_3 x + b_3 y + c_3 z &= d_3 \end{aligned} \tag{5}$$

yields the fractions which have the denominator of the form

$$a_1b_2c_3 - a_1c_2b_3 - b_1a_2c_3 + b_1c_2a_3 + c_1a_2b_3 - c_1b_2a_3 \qquad (6)$$

The last expression is called the determinant of the third order and is designated as

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} \qquad (7)$$

Transforming expression (6) and bearing in mind notation (3) we derive the formula

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1(b_2c_3 - c_2b_3) - b_1(a_2c_3 - c_2a_3) +$$

$$+ c_1(a_2b_3 - b_2a_3) = a_1\begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - b_1\begin{vmatrix} a_2 & c_2 \\ a_3 & c_3 \end{vmatrix} + c_1\begin{vmatrix} a_2 & b_2 \\ a_3 & b_3 \end{vmatrix} \qquad (8)$$

which is of use for calculating a determinant. For instance,

$$\begin{vmatrix} 1 & 0 & -2 \\ -1 & 1 & \frac{1}{2} \\ 3 & 1 & 2 \end{vmatrix} = 1\begin{vmatrix} 1 & \frac{1}{2} \\ 1 & 2 \end{vmatrix} - 0\begin{vmatrix} -1 & \frac{1}{2} \\ 3 & 2 \end{vmatrix} + (-2)\begin{vmatrix} -1 & 1 \\ 3 & 1 \end{vmatrix} =$$

$$= 1\left(1\cdot 2 - \frac{1}{2}\cdot 1\right) - 2(-1\cdot 1 - 1\cdot 3) = \frac{3}{2} + 8 = 9\frac{1}{2}$$

Determinants of the fourth order are defined by analogy with formula (8):

$$\begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{vmatrix} = a_1\begin{vmatrix} b_2 & c_2 & d_2 \\ b_3 & c_3 & d_3 \\ b_4 & c_4 & d_4 \end{vmatrix} - b_1\begin{vmatrix} a_2 & c_2 & d_2 \\ a_3 & c_3 & d_3 \\ a_4 & c_4 & d_4 \end{vmatrix} +$$

$$+ c_1\begin{vmatrix} a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \\ a_4 & b_4 & d_4 \end{vmatrix} - d_1\begin{vmatrix} a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \\ a_4 & b_4 & c_4 \end{vmatrix}$$

(we suggest that the reader should carefully think over the structure of the expression entering in the right-hand side). Determinants of the fifth, sixth etc. order (and, generally, determinants of the nth order) are introduced in a similar way.

2. Properties. We are going to describe the properties of determinants but we shall take only the determinants of the third order of form (7) although all these properties hold for determinants of any order.

A determinant of the third order of form (7) has three *rows* and three *columns*. It consists of nine *elements* (i.e. of the numbers $a_1, b_1, \ldots, c_3$).

1. Interchanging two rows or two columns of a determinant is equivalent to multiplying the determinant by -1. For example,

$$
\begin{vmatrix}
a_1 & b_1 & c_1 \\
a_2 & b_2 & c_2 \\
a_3 & b_3 & c_3
\end{vmatrix}
= -
\begin{vmatrix}
c_1 & b_1 & a_1 \\
c_2 & b_2 & a_2 \\
c_3 & b_3 & a_3
\end{vmatrix}
\tag{9}
$$

(we have interchanged the third and the first columns). This is proved by comparing both sides of (9) according to formula (8):

$$
-
\begin{vmatrix}
c_1 & b_1 & a_1 \\
c_2 & b_2 & a_2 \\
c_3 & b_3 & a_3
\end{vmatrix}
= -c_1
\begin{vmatrix}
b_2 & a_2 \\
b_3 & a_3
\end{vmatrix}
+ b_1
\begin{vmatrix}
c_2 & a_2 \\
c_3 & a_3
\end{vmatrix}
- a_1
\begin{vmatrix}
c_2 & b_2 \\
c_3 & b_3
\end{vmatrix}
=
$$

$$
= -c_1 (b_2 a_3 - a_2 b_3) - b_1 (c_2 a_3 - a_2 c_3) - a_1 (c_2 b_3 - b_2 c_3)
$$

If we remove the parentheses here we shall obtain the expression equal to (6). This proves formula (9).

2. A determinant having two identical rows or columns is equal to zero. For example,

$$
P =
\begin{vmatrix}
a_1 & b_1 & c_1 \\
a_2 & b_2 & c_2 \\
a_2 & b_2 & c_2
\end{vmatrix}
= 0
$$

(here the second row coincides with the third one). Virtually, if we interchange the two rows then by property 1 we get $-P = P$, i.e. $P = 0$.

3. A common factor entering into all the elements of a row or of a column can be taken outside the determinant. For instance,

$$
\begin{vmatrix}
a_1 & k b_1 & c_1 \\
a_2 & k b_2 & c_2 \\
a_3 & k b_3 & c_3
\end{vmatrix}
= k
\begin{vmatrix}
a_1 & b_1 & c_1 \\
a_2 & b_2 & c_2 \\
a_3 & b_3 & c_3
\end{vmatrix}
\tag{10}
$$

The proof is carried out by verifying the equality.

4. A determinant having a row or a column consisting of zeros equals zero. In order to prove this assertion it is sufficient to put $k = 0$ in formula (10).

5. If each element of a row or of a column (for instance. of the second row) can be represented in the form of a sum of two terms the determinant itself can be represented as a sum of two determi-

nants according to the formula

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2'+a_2'' & b_2'+b_2'' & c_2'+c_2'' \\ a_3 & b_3 & c_3 \end{vmatrix} = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2' & b_2' & c_2' \\ a_3 & b_3 & c_3 \end{vmatrix} + \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2'' & b_2'' & c_2'' \\ a_3 & b_3 & c_3 \end{vmatrix}$$

The proof is carried out by verifying the equality of both sides.

6. Adding arbitrary numbers proportional to the elements of a row (a column) to the corresponding element of another row (column) of a determinant we do not change the numerical value of the determinant. Indeed, for instance,

$$\begin{vmatrix} a_1+kc_1 & b_1 & c_1 \\ a_2+kc_2 & b_2 & c_2 \\ a_3+kc_3 & b_3 & c_3 \end{vmatrix} = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} + \begin{vmatrix} kc_1 & b_1 & c_1 \\ kc_2 & b_2 & c_2 \\ kc_3 & b_3 & c_3 \end{vmatrix} =$$

$$= \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} + k \begin{vmatrix} c_1 & b_1 & c_1 \\ c_2 & b_2 & c_2 \\ c_3 & b_3 & c_3 \end{vmatrix} = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}$$

(in the calculations we have applied, in succession, properties 5, 3 and 2).

7. The value of a determinant does not change if each of its rows is replaced by the corresponding columns and vice versa, that is

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}$$

(this is the operation of **transposing a determinant**). The proof is carried out by verifying the equality.

3. **Expanding a Determinant in Minors of Its Row or Column.** First we introduce the notion of a cofactor of an element of a determinant. Suppose we choose an element of a determinant and then delete the row and the column to which the element belongs. Thus we get a determinant of lower order which is called the **minor** (**minor determinant**) of the element of a determinant. Now let us supply each minor with the sign + or − depending on the position the corresponding element occupies in the original determinant according to the following rule: we take + for the minor of the element standing in the left upper corner of a determinant (this element is sometimes called the origin of the determinant) and alternate the signs of other elements in chess-board order according to the scheme

$$\begin{array}{ccc} + & - & + \\ - & + & - \\ + & - & + \end{array}$$

The quantities thus obtained are called **cofactors** (or **signed minors** or **algebraic adjuncts**) of the elements of a determinant. For instance, the cofactor A_1 of the element a_1 [see determinant (7)] is equal to $\begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix}$, the cofactor C_2 of the element c_2 is equal to $-\begin{vmatrix} a_1 & b_1 \\ a_3 & b_3 \end{vmatrix}$ etc.

There is a general property of determinants which is formulated as follows: *a determinant is equal to the sum of the products of the elements of any row or column of the determinant by their cofactors.* For example,

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = b_1 B_1 + b_2 B_2 + b_3 B_3 =$$

$$= -b_1 \begin{vmatrix} a_2 & c_2 \\ a_3 & c_3 \end{vmatrix} + b_2 \begin{vmatrix} a_1 & c_1 \\ a_3 & c_3 \end{vmatrix} - b_3 \begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}$$

This representation of a determinant is called the *expansion of the determinant in minors of its row or column* (we have the expansion in minors of the second column in the example). We also say that the determinant is expanded in terms of the elements of its row or column. As in Sec. 2, the proof is based on verifying the equality.

The properties enumerated in Secs. 2 and 3 are applied to evaluating determinants. For example, let us compute the determinant

$$D = \begin{vmatrix} 1 & 0 & 2 & -1 \\ 3 & 1 & 0 & -1 \\ 2 & 1 & -1 & 0 \\ 0 & 3 & 2 & 1 \end{vmatrix}$$

Here we can apply property 6 of Sec. 2 and make all the elements of a certain row or column but one be zero. Then expanding the determinant in minors of this row or column we get only one nonzero summand because all other minors are multiplied by zeros. If, for instance, we want only the element occupying the second place in the third row of the determinant to be unequal to zero we should multiply the second column by -2, add it to the first column and after this add the second column of the resulting determinant to its third column. Then we obtain

$$D = \begin{vmatrix} 1 & 0 & 2 & -1 \\ 1 & 1 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ -6 & 3 & 2 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 2 & -1 \\ 1 & 1 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ -6 & 3 & 5 & 1 \end{vmatrix}$$

(the two operations are usually carried out simultaneously, by "one step"). Now expanding in minors of the third row we obtain

$$D = -1 \cdot \begin{vmatrix} 1 & 2 & -1 \\ 1 & 1 & -1 \\ -6 & 5 & 1 \end{vmatrix}$$

Here we can, for example, subtract the second row from the first one which results in

$$D = -\begin{vmatrix} 0 & 1 & 0 \\ 1 & 1 & -1 \\ -6 & 5 & 1 \end{vmatrix}$$

Now if we expand D in minors of the first row we finally obtain

$$D = -(-1) \begin{vmatrix} 1 & -1 \\ -6 & 1 \end{vmatrix} = 1 \cdot 1 - (-1)(-6) = -5$$

In case the elements of a determinant are approximate numbers the method is used in a modified form. We shall illustrate this technique by evaluating the determinant

$$D = \begin{vmatrix} -1.37 & 2.15 & 0.76 \\ 2.31 & -1.78 & -4.32 \\ -0.86 & 2.13 & 3.25 \end{vmatrix}$$

Let us factor out the first element of the first row using the rule of a reserve decimal digit (see Sec. I.9):

$$D = -1.37 \begin{vmatrix} 1 & -1.569 & -0.555 \\ 2.31 & -1.78 & -4.32 \\ -0.86 & 2.13 & 3.25 \end{vmatrix}$$

Multiply the first row by -2.31 and add it to the second row and, simultaneously, multiply the first row by 0.86 and add it to the third row:

$$D = -1.37 \begin{vmatrix} 1 & -1.569 & -0.555 \\ 0 & 1.844 & -3.038 \\ 0 & 0.781 & 2.773 \end{vmatrix} = -1.37 \begin{vmatrix} 1.844 & -3.038 \\ 0.781 & 2.773 \end{vmatrix}$$

Let us repeat this procedure:

$$D = -1.37 \times 1.844 \begin{vmatrix} 1 & -1.647 \\ 0.781 & 2.773 \end{vmatrix} =$$

$$= 1.37 \times 1.844 \begin{vmatrix} 1 & -1.647 \\ 0 & 4.059 \end{vmatrix} = -1.37 \times 1.844 \times 4.059 = -10.3$$

This method is also applicable to determinants of higher order. It cannot be used in case some of the determinants occurring in the process of calculating contain the number 0 (or a number which is close to zero and known with a low degree of relative accuracy) as its left uppermost element. To overcome the difficulty we can slightly change the method and begin not with the left uppermost element but with the element which is the greatest in its absolute value. This element (the principal element) should be factored out of the row (column) in which it is contained. (The element -4.32 is the principal element in our previous example.)

The fundamentals of the theory of determinants were introduced in 1750 by the Swiss mathematician G. Cramer (1704-1752).

§ 2. Systems of Linear Algebraic Equations

4. Basic Case. We shall limit ourselves to such systems in which the number of equations coincides with the number of the unknowns. We shall deal only with the systems of three equations, i.e. systems of type (5). For example, if we want to find y we should multiply the first of equations (5) by the cofactor B_1 of the element b_1 of determinant (7), multiply the second equation by B_2 and the third one by B_3 and then sum together the results. Doing this we derive

$$(a_1B_1 + a_2B_2 + a_3B_3)\, x + (b_1B_1 + b_2B_2 + b_3B_3)\, y +$$
$$+ (c_1B_1 + c_2B_2 + c_3B_3)\, z = d_1B_1 + d_2B_2 + d_3B_3 \qquad (11)$$

But the expression inside the first parentheses is equal to

$$\begin{vmatrix} a_1 & a_1 & c_1 \\ a_2 & a_2 & c_2 \\ a_3 & a_3 & c_3 \end{vmatrix} \qquad (12)$$

In fact, if we expand the determinant in minors of the second column we obtain the sum of the products of the elements a_1, a_2 and a_3 by their cofactors, these cofactors in determinant (12) being equal to the cofactors of the elements b_1, b_2 and b_3 in determinant (7), i.e. equal to B_1, B_2 and B_3, respectively. Determinant (12), by property 2 in Sec. 2, is therefore equal to zero. By the same reason, the expression in the third parentheses of formula (11) also equals zero whereas the right-hand side of (11) is equal to

$$\begin{vmatrix} a_1 & d_1 & c_1 \\ a_2 & d_2 & c_2 \\ a_3 & d_3 & c_3 \end{vmatrix}$$

The expression in the second parentheses in (11) is just equal to determinant (7) itself by Sec. 3. This determinant consists of the coefficients in unknown quantities in system (5) and is called the

determinant of the system; we shall denote it, for brevity, by the letter D. Thus, from (11) we deduce

$$Dy = \begin{vmatrix} a_1 & d_1 & c_1 \\ a_2 & d_2 & c_2 \\ a_3 & d_3 & c_3 \end{vmatrix} \tag{13}$$

In the same way we find

$$Dx = \begin{vmatrix} d_1 & b_1 & c_1 \\ d_2 & b_2 & c_2 \\ d_3 & b_3 & c_3 \end{vmatrix} \quad \text{and} \quad Dz = \begin{vmatrix} a_1 & b_1 & d_1 \\ a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \end{vmatrix} \tag{14}$$

Now let us first suppose that $D \ne 0$, this is the basic case. As it will be shown in the end of Sec. X.7, in this case the system in question has a unique solution. From (13) and (14) we derive the formulas for the solution:

$$x = \frac{\begin{vmatrix} d_1 & b_1 & c_1 \\ d_2 & b_2 & c_2 \\ d_3 & b_3 & c_3 \end{vmatrix}}{D}, \quad y = \frac{\begin{vmatrix} a_1 & d_1 & c_1 \\ a_2 & d_2 & c_2 \\ a_3 & d_3 & c_3 \end{vmatrix}}{D}, \quad z = \frac{\begin{vmatrix} a_1 & b_1 & d_1 \\ a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \end{vmatrix}}{D}$$

Thus, *every unknown quantity is equal to a quotient which has the determinant of the system in the denominator, the numerator being the determinant which is obtained from the determinant of the system by substituting the column consisting of the right-hand members of system* (5) *for the column of the coefficients in that unknown.*
Let us take, for example, the system of equations

$$\left.\begin{array}{r} x - 2z = 1 \\ 2x + y - z = 0 \\ x - 2y + z = -2 \end{array}\right\}$$

Suppose it is necessary to find the value of z. Then

$$z = \frac{\begin{vmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ 1 & -2 & -2 \end{vmatrix}}{\begin{vmatrix} 1 & 0 & -2 \\ 2 & 1 & -1 \\ 1 & -2 & 1 \end{vmatrix}} = \frac{1(-2+0)-0+1(-4-1)}{1(1-2)-0-2(-4-1)} = \frac{-2-5}{-1+10} = -\frac{7}{9}$$

In the same way we compute the remaining unknowns and it is only the numerators that should be evaluated since the denominators equal $D = 9$.

The above formulas are called **Cramer's rule.** If we apply the formulas to system (1) of two equations in two unknowns we obtain formulas (4).

5. Numerical Solution. The application of the formulas given in Sec. 4 is inconvenient in case the coefficients of equations are approximate numbers and, in particular, if the number of equations is large.

There are many methods that can be used in such cases. Here we shall represent two of them. We again take system (5) as an example but the methods are applicable to systems with any arbitrary number of unknowns.

Gauss' method (the method of elimination) named after K. F. Gauss a famous German mathematician (1777-1855), who also obtained fundamental results in astronomy and geodesy, is analogous to the method of evaluating determinants in the end of Sec. 3.

We divide the first equation of system (5) by a_1 which results in

$$x + b_1'y + c_1'z = d_1' \tag{15}$$

(the primes designate the new coefficients here). Multiply equation (15) by $-a_2$ $(-a_3)$ and add it to the second (third) equation of the system. This eliminates x and yields the system

$$\left. \begin{array}{l} b_2'y + c_2'z = d_2' \\ b_3'y + c_3'z = d_3' \end{array} \right\} \tag{16}$$

Now we divide the first of equations (16) by b_2' and obtain

$$y + c_2''z = d_2'' \tag{17}$$

Multiplying the last equation by $-b_3'$ and adding it to the second of equations (16) we derive the equation of the form

$$c_3''z = d_3'' \tag{18}$$

that is y is also eliminated. From (18) we find z; substituting it into (17) we determine y; then substituting y and z into (15) we finally evaluate x.

As in Sec. 3, if at some stage of the calculations the left uppermost coefficient turns out to be considerably smaller than others in its absolute value the method should be modified, that is we should eliminate the unknown whose coefficient is the greatest in the absolute value by dividing the corresponding equation by the coefficient.

The **iterative method** (compare with Sec. V.3) is applied to system (5) in the following way. The system is rewritten in the form

$$\left. \begin{array}{l} x = \alpha_1 x + \beta_1 y + \gamma_1 z + \delta_1 \\ y = \alpha_2 x + \beta_2 y + \gamma_2 z + \delta_2 \\ z = \alpha_3 x + \beta_3 y + \gamma_3 z + \delta_3 \end{array} \right\} \tag{19}$$

Then we choose certain values $x = x_0$, $y = y_0$ and $z = z_0$ as the *zeroth approximation*. The values are substituted for x, y and z,

respectively, into the right-hand sides of system (19) and thus the *first approximation* x_1, y_1 and z_1 is obtained. Substituting again x_1, y_1 and z_1 into the right-hand sides of (19) we get the *second approximation* and so forth. Generally, the $(n + 1)$th approximation is expressed in terms of the nth approximation by the formulas

$$\left.\begin{array}{l} x_{n+1} = \alpha_1 x_n + \beta_1 y_n + \gamma_1 z_n + \delta_1 \\ y_{n+1} = \alpha_2 x_n + \beta_2 y_n + \gamma_2 z_n + \delta_2 \\ z_{n+1} = \alpha_3 x_n + \beta_3 y_n + \gamma_3 z_n + \delta_3 \end{array}\right\} \tag{20}$$

If the process converges, that is the successive approximations have limits as $n \to \infty$ ($x_n \to \bar{x}$, $y_n \to \bar{y}$ and $z_n \to \bar{z}$), then passing to the limit in formulas (20) as $n \to \infty$ we see that $x = \bar{x}$, $y = \bar{y}$ and $z = \bar{z}$ form the solution of system (19).

Just as in the examples given in Sec. V.3, the smaller the absolute values of the coefficients in the unknowns in the right-hand sides of the system, the faster the convergence of the iterative method for system (19). Some more precise tests for the convergence will be given in Sec. XVII.18.

System (5) is solvable in the simplest way in case it has the *triangular form*:

$$\left.\begin{array}{l} a_1 x \qquad\qquad\quad = d_1 \\ a_2 x + b_2 y \qquad = d_2 \\ a_3 x + b_3 y + c_3 z = d_3 \end{array}\right\}$$

In this case we immediately find x from the first equation of the system; substituting it into the second equation we obtain y and then determine z after substituting x and y into the third equation. Gauss' method described above is essentially the method of reducing general system (5) to the triangular form (15), (17) and (18).

Different useful rules for numerical solution of systems of linear algebraic equations can be found in [3], [10] and [13].

6. Singular Case. If the determinant of a system is equal to zero we shall say that there is a singular case. As it will be shown in Sec. X.7, in such a case there may be one of the following two possibilities.

1. The system is inconsistent (that is contradictory) which means that it has no solution. For example, the system

$$\left.\begin{array}{l} x + 2y - z = 1 \\ 2x - y = 3 \\ 3x + y - z = 5 \end{array}\right\}$$

is just of this type since adding together the first two equations we arrive at a contradiction with the third one.

Vectors

§ 1. Linear Operations on Vectors

1. Scalar and Vector Quantities. Scalar and vector quantities differ in the following aspect: whilst the former are completely characterized by their numerical values relative to a chosen system of measurement units (such quantities as temperature, work, density etc.), the latter have, in addition, certain direction in space (such quantities as force, velocity etc.). All the quantities we studied before were scalars and it was therefore permissible not to use the word "scalar". But when we consider both scalar and vector quantities it is essentially important to take into account the nature of the quantities. A vector quantity, or, simply, a **vector**, can be represented by a line segment in space if we choose a certain unit of length. For instance, if we intend to represent forces we can assume that the segment of 2 cm represents the force of 1 kg and the like (see Sec. I.4 on this question). The line segment representing a vector is directed (oriented), that is its **origin** and its **terminus** must be indicated. The direction of a vector is usually indicated by an arrow. Diverting a vector from its direction we obtain the absolute value **(modulus)** of this quantity. Thus, the absolute value of a vector is a scalar which has a dimension of the quantity in question and is always positive with the only exception for the zero vector (see Sec. 3). The absolute value of a vector is also called its length. For example, let us take a vector representing a force. When we speak about the vector we mean that its length has the dimension of force, that is the measurement of the length of the representing line segment in the chosen scale units yields the magnitude of the force. In mathematics vectors are usually regarded as dimensionless. The absolute value of such a vector is a dimensionless (abstract) number. Vectors are usually given in bold face or designated by arrows (see Fig. 139). The absolute value of a vector is denoted by the same letter but in ordinary print or in bold face with vertical bars (the sign of modulus): $AB = a = |\mathbf{a}|$.

Thus, to define a vector means to define its absolute value and its direction in space. Accordingly, every vector can be translated (that is it can be transferred in such a way that its direction in space should remain parallel to the original direction) to any place. Hence, the origin of a vector ("point of application") can be anywhere. In other words, *two vectors are regarded as being equal if they have the same absolute values, are parallel and similarly directed* (see Fig. 140).

The freedom of translating a vector is sometimes restricted. For instance, the origin of a certain vector can be fixed. Such vectors are called *localized* or *bound vectors* (a radius-vector mentioned in
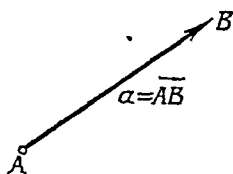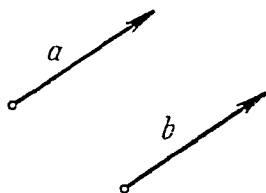
Fig. 139   Fig. 140
                                    a=b

Sec. 9 presents an example of such a vector). Further, there can exist a certain straight line in which a vector must lie. Then the vector is said to be a *sliding vector*. The vector of angular velocity of a rotary motion which lies on the axis of revolution is an example of such a vector. If the parallel translation of a vector is unrestricted the vector is called a *free vector*.

2. Addition of Vectors. Linear operations on vectors are the operations of adding vectors together and of multiplying a vector by a scalar (and the operation of subtraction which is, of course, connected with the addition). Generally, a quantity can be considered to be a vector if and only if the above operations are performed in accordance with the rules described further in Secs. 2-4.

The addition of two vectors is performed according to the well-known *parallelogram law* in mechanics which is the rule of adding forces and velocities. For example, if it is required to add together two vectors a and b they are applied to a common origin and then a parallelogram is constructed on them (see Fig. 141). A vector coinciding with the diagonal of the parallelogram whose origin is that of a and b is, by definition, the sum a + b. This construction straightway implies that a + b = b + a, that is the commutative law holds for the addition of vectors.

The opposite sides of a parallelogram being parallel and equal, the vector $\overrightarrow{DC}$ in Fig. 141 is also equal to b. This implies one more rule of adding vectors: the origin of the second vector is placed

at the terminus of the first vector. Then a vector closing the triangle, that is the vector whose origin is that of the first vector and whose terminus is that of the second vector, is the sum of the vectors (Fig. 142). If now it is necessary to add a third vector to the above sum we must put the origin of the third vector at the terminus of
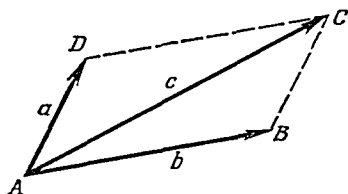
Fig. 141
c=a+b
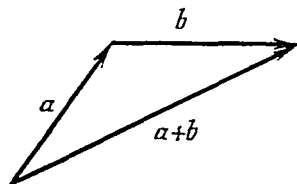
Fig. 142

the second one and take the closing vector again etc. The general rule of adding together any number of vectors is illustrated in Fig. 143. It follows from Fig. 144 that the **associative law** also holds for a sum of vectors: $a + (b + c) = (a + b) + c$. The commutative and associative laws imply that the order of summation and the way of bracketing do not affect a sum of any number of summands. For example,

$$(a + b) + (c + d) = [(b + d) + c] + a =$$
$$= [c + (a + d)] + b$$

and the like.

We must underline that we cannot add together vectors of different dimensions and that it is impossible to add together a vector
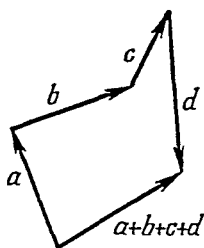
Fig. 143
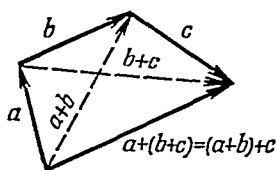
Fig. 144

and a scalar. Besides, we shall not consider the comparison of vectors in our course. This means that there will be no positive and negative vectors or inequalities of the form $a > b$ etc. But of course we can compare absolute values (lengths) of vectors. At the same time we must not be surprised that the absolute value of a sum of vectors may happen to be, for example, less than the absolute value

of each summand. Actually, vectors are added not as numbers but according to the parallelogram law of forces, and the resultant of a system of forces can be smaller than each of the forces.

We conclude by pointing out that there is a consequence of Fig. 143, namely,

$$| a + b + c + d | \leqslant | a | + | b | + | c | + | d |$$

The equality will be here only if all the vectorial addends are of the same direction; in such a case the polygon of vectors degenerates into a straight line.

3. **Zero Vector and Subtraction of Vectors.** A vector whose terminus coincides with its origin is called the zero vector or, simply, zero. Its absolute value is equal to zero whereas all the other vectors have positive absolute values. The direction of this vector is
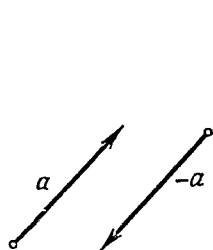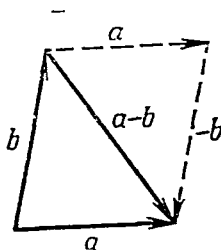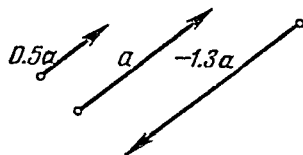
Fig. 145   Fig. 146   Fig. 147

undetermined, i.e. any direction may be ascribed to it. We can therefore regard the zero vector as parallel (perpendicular) to any vector. It is denoted as 0 and its role in an operation of adding vectors is similar to that of the number zero in adding numbers. In fact, it is apparent that $a + 0 = a$.

Let a vector $a = \overrightarrow{AB}$ be given. Then the vector $\overrightarrow{BA}$ is called the negative of the vector a and is denoted as $-a$ (see Fig. 145). Obviously, $a + (-a) = 0$.

To subtract a vector means to add its negative. It follows that

$$b + (a - b) = b + [a + (-b)] = a + [b + (-b)] =$$
$$a + 0 = a$$

This corresponds to the usual definition of a difference. The geometrical interpretation of the rule of subtraction is shown in Fig. 146.

4. **Multiplying a Vector by a Scalar.** The product $\lambda a = a\lambda$ of a vector a by a dimensionless scalar (number) λ is defined as follows: if $\lambda > 0$ then the product is a vector which is obtained from a as it is stretched λ-fold without changing its direction; if $\lambda < 0$ then a must be stretched $| \lambda |$-fold and its direction must be replaced

by the opposite direction (see Fig. 147). Further, $\frac{a}{\lambda} = \left(\frac{1}{\lambda}\right)$ a. These definitions imply the following simple properties:

1. (-1) a $= -$a;
2. 0a $= 0$;
3. $\lambda 0 = 0$;
4. $(\lambda + \mu)$ a $= \lambda$a $+ \mu$a;
5. λ (a $+$ b) $= \lambda$a $+ \lambda$b;
6. λ (μa) $= (\lambda\mu)$ a;
7. $\lambda \frac{a}{\lambda} = $ a;
8. If n is a positive integer then na $= \underbrace{a + a + \ldots + a}_{n \text{ times}}$.

These properties enable us to perform linear operations on vectors and transform algebraic expressions containing vectors in the same way as it is done with numbers. The properties are proved in an obvious way. For instance, Fig. 148 demonstrates the proof of property 5 $[\lambda a + \lambda b = \lambda c = \lambda (a + b)]$ for $\lambda > 0$.

If a scalar λ has a dimension the product λa is defined as a vector whose absolute value is equal to $|\lambda| |$ a $|$ and which is parallel to a and directed like a if $\lambda > 0$ and directed contrary to a if $\lambda < 0$.
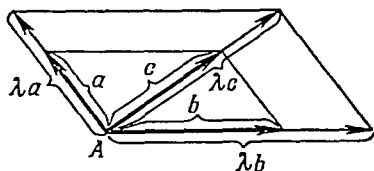
Fig. 148
$c = a - b, \quad \lambda c = \lambda a - \lambda b$
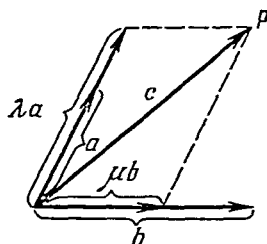
Fig. 149
$c = \lambda a + \mu b$

All the enumerated properties remain true for this case as well. Further, for the sake of simplicity, we shall regard all vectors and scalars as dimensionless unless otherwise stated.

5. **Linear Combination of Vectors.** Suppose we have several vectors, for example, three vectors a, b and c. Then every vector of the form d $= \lambda$a $+ \mu$b $+ \nu$c where λ, μ and ν are scalars is called a **linear combination** of the vectors a, b and c. The vector d is also said to be *expressed linearly* in terms of a, b and c which means that d can be obtained by linear operations on a, b and c. As examples of such linear combinations we can take

$$a + 2b - 3c, \quad \frac{a+c}{2} = \frac{1}{2} a + 0b + \frac{1}{2} c, \quad 0 = 0 \cdot a + 0 \cdot b + 0 \cdot c \text{ etc.}$$

A given system of vectors is called **linearly dependent** if one of the vectors is expressed linearly in terms of the rest. If otherwise the vectors are called **linearly independent**.

Two vectors are linearly dependent if and only if they are parallel to each other. Certainly, it follows from the definition (see Sec. 4) that $b = \lambda a$ implies $b \parallel a$. Conversely, if two vectors are parallel then they are linearly dependent because we can always take such a coefficient of extension λ that after increasing the length of one of the vectors λ times the extended vector should coincide with the other vector (in case the parallel vectors are of opposite directions the coefficient λ is negative).

Three vectors are linearly dependent if and only if they are parallel to a common plane. Virtually, let $c = \lambda a + \mu b$. Let us translate the three vectors so that their origins coincide. Draw a plane through the vectors a and b (the plane P in Fig. 149). Then the vectors λa and μb will lie in the plane P and therefore their sum, that is c, will also lie in the same plane. Consequently, the vectors a, b and c were parallel to the plane P in their original positions. Conversely, let it be known that the vectors a, b and c are parallel to a common plane P. Fig. 149 shows the representation of c in the form of linear combination of a and b in case $a \nparallel b$. Such a representation is called *the resolution of a vector in a plane into components with respect to two given non-parallel vectors.* As for the case $a \parallel b$, the preceding paragraph implies that one of the two vectors a and b (for instance, the vector a) is expressed linearly in terms of the other vector (for instance, in terms of b). Therefore a, b and c are linearly dependent because a is expressed in terms of b and c.

Four or more vectors are always linearly dependent. Indeed, let us take four vectors a, b, c and d. Translate them to a common origin. If after this the vectors a, b and c lie in a common plane then, in accord with the preceding paragraph, one of them is expressed linearly in terms of the rest etc. (as in the end of the preceding paragraph). Let now a, b and c not lie in a common plane after they are translated to a common origin (see Fig. 150). Then we can draw a straight line parallel to the vector c and passing through the point D (the terminus of the vector d), the point C being the point of intersection of the straight line with the plane in which the vectors a and b lie. Finally, we draw a straight line parallel to the vector b and passing through the point C. This straight line intersects the straight line containing the vector a at a point B. Then we can write

$$d = \overrightarrow{AD} = \overrightarrow{AB} + \overrightarrow{BC} + \overrightarrow{CD} = \lambda a + \mu b + \nu c \qquad (1)$$

Such a representation is called *the resolution of a vector into components with respect to three given vectors which are not parallel to one and the same plane.* We also call it *the resolution of a vector into*

components along three given axes (these axes are denoted as *ll*, *mm* and *nn* in Fig. 150). Representations of this type are utilized in theoretical mechanics and other branches of science for resolving
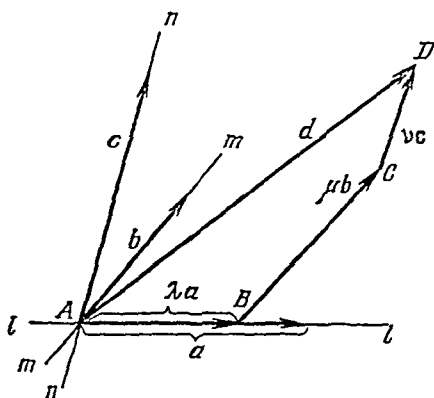
Fig. 150
$$d = \lambda a + \mu b + \nu c$$

forces and other vector quantities into components along three given directions. Each of the summands λa, μb and νc is called a
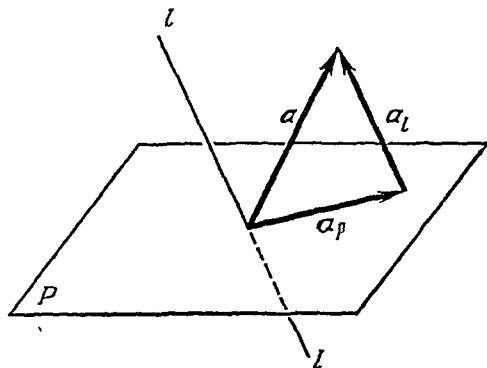
Fig. 151
$$a = a_p + a_l$$

component of the vector d along the corresponding direction. A component along an axis depends not only on the direction of the axis but also on the directions of other axes. At the same time a component does not depend on the choice of positive directions for given axes.

Another type of representing a vector which is sometimes used is shown in Fig. 151. Here a given vector is resolved into components a_p and a_l where a_p lies in a given plane and a_l is parallel to a given axis.

Resolution (1) can be performed uniquely. Indeed, if another representation $d = \lambda_1 a + \mu_1 b + \nu_1 c$ existed we should equate their right-hand sides and deduce the relation $(\lambda - \lambda_1) a + (\mu - \mu_1) b + (\nu - \nu_1) c = 0$. This would imply that the vectors a, b and c would be linearly dependent (why is it so?).

A system of linearly independent vectors used for resolving other vectors into components with respect to the system is called a basis. The foregoing discussion implies that any two non-parallel vectors can be taken as a basis in their plane and that any three vectors non-parallel to one and the same plane can be taken as a basis in space. If a, b and c form such a basis then the numbers λ, μ and ν entering into representation (1) are called the **coordinates** of the vector d with respect to the basis a, b, c. If a basis a, b, c is given the coordinates λ, μ and ν are uniquely determined by the vector d. Conversely, the vector d is uniquely determined by its coordinates λ, μ and ν.

§ 2. Scalar Product of Vectors

6. Projection of Vector on Axis. Let a vector $a = \overrightarrow{AB}$ and an axis l be given (see Fig. 152). The projection $\text{proj}_l\, a$ of the vector a on the axis l is the length of the segment $A'B'$ connecting the feet
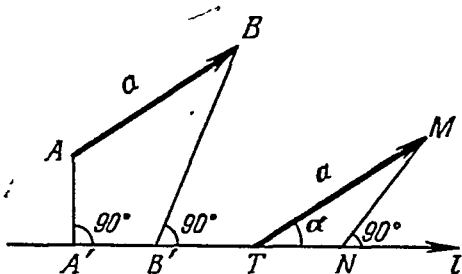
Fig. 152
Note that the drawing is spatial!

of the perpendiculars drawn from the points A and B to the axis l. This length is taken with the sign $+$ or $-$ depending on whether the direction of the segment $A'B'$ coincides with the positive direction of the axis or is opposite to it. A projection of a vector on another vector is defined similarly. In this case the perpendiculars are drawn to the other vector or to its prolongation. Thus, *a projection of a vector is a scalar*; the dimension of a projection coincides with that of the projected vector.

components along three given axes (these axes are denoted as *ll*, *mm* and *nn* in Fig. 150). Representations of this type are utilized in theoretical mechanics and other branches of science for resolving

Fig. 150
$$d = \lambda a + \mu b + vc$$

forces and other vector quantities into components along three given directions. Each of the summands λa, μb and vc is called a
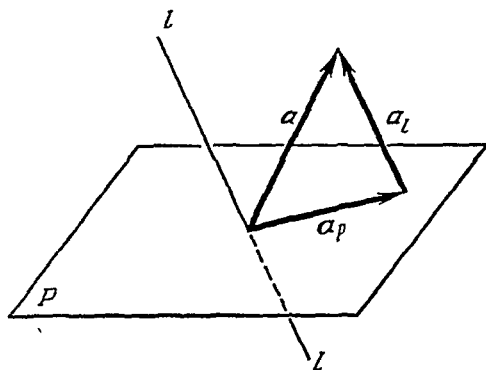
Fig. 151
$$a = a_p + a_l$$

component of the vector **d** along the corresponding direction. A component along an axis depends not only on the direction of the axis but also on the directions of other axes. At the same time a component does not depend on the choice of positive directions for given axes.

formula (3) cannot be extended to more than two vectors. Bearing in mind formula (2) we can also put down the relation

$$\mathbf{a} \cdot \mathbf{b} = b \, \text{proj}_b \, \mathbf{a} = a \, \text{proj}_a \, \mathbf{b} \tag{4}$$

Therefore, *the scalar product of two vectors is equal to the product of the absolute value of one of the vectors by the projection of the other vector on the first.*

Example. Let s be a displacement vector of a material point and let F be a constant force (one of the forces acting upon the point in a process of motion) depicted in Fig. 155. To reckon the work A performed by this force we must take into account only the component $\mathbf{F}'$ of the force F along the direction of the displacement. Hence, $A = s \, \text{proj}_s \, \mathbf{F} = \mathbf{s} \cdot \mathbf{F}$.

The dimension of a scalar product is equal to the product of dimensions of the factors.
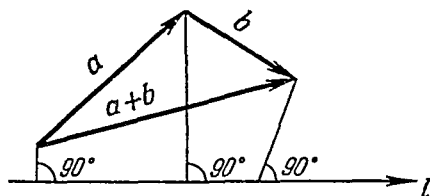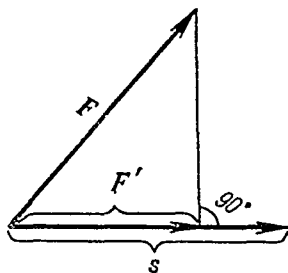
Fig. 154   Fig. 155

Formulas (3) and (4) are simplified if one of the factors (or both factors) is a **unit vector**, that is a vector with the dimensionless absolute value 1. For example, if $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}$ are unit vectors,

$$\mathbf{e}_1 \cdot \mathbf{e}_2 = \cos (\widehat{\mathbf{e}_1, \, \mathbf{e}_2}), \quad \mathbf{a} \cdot \mathbf{e} = \text{proj}_e \, \mathbf{a} \tag{5}$$

A unit vector lying on an axis l is usually denoted as $\mathbf{l}^\circ$. Similarly, a unit vector parallel to a vector **b** and having the same direction is denoted by $\mathbf{b}^\circ$. Taking into account formula (5) we can write

$$\text{proj}_l \, \mathbf{a} = \text{proj}_{l^\circ} \, \mathbf{a} = \mathbf{a} \cdot \mathbf{l}^\circ, \quad \text{proj}_b \, \mathbf{a} = \mathbf{a} \cdot \mathbf{b}^\circ$$

8. Properties of Scalar Product.

1. A scalar product is equal to zero if and only if the vectors are **orthogonal**, that is perpendicular to each other:

$$\mathbf{a} \cdot \mathbf{b} = 0 \text{ is equivalent to } \mathbf{a} \perp \mathbf{b}$$

Virtually, this follows from (3) since $\cos (\widehat{a, \, b}) = 0$ implies $(\widehat{a, \, b}) = 90°$. Of course, it may happen that $a = 0$ but this means

that $a = 0$ and a zero vector can be regarded as being perpendicular to any vector (see Sec. 3).

2. $a \cdot a = a^2$ because $(\widehat{a, a}) = 0°$ and $\cos (\widehat{a, a}) = 1$. In other words, the scalar square of a vector equals the square of its absolute value.

3. A scalar product does not depend on the order of its factors: $a \cdot b = b \cdot a$. [This immediately follows from formula (3).]

4. A scalar factor may be taken outside the scalar product of vectors:

$$(\lambda a) \cdot b = a \cdot (\lambda b) = \lambda (a \cdot b) \tag{6}$$

Really, on the basis of property 4 from Sec. 6, we have $\text{proj}_b (\lambda a) = \lambda \text{proj}_b a$. Multiplying both sides by b and taking into account formula (4) we derive $b \cdot (\lambda a) = \lambda (a \cdot b)$. The scalar product being commutative, the last relation implies formula (6).

[We suggest that the reader should deduce formula (6) as a direct consequence of the definition of a scalar product.]

5. Distributive law:

$$(a + b) \cdot c = a \cdot c + b \cdot c$$

To prove this property we write $\text{proj}_c (a + b) = \text{proj}_c a + \text{proj}_c b$ on the basis of property 5 in Sec. 6. Then we multiply both sides of the last relation by c which yields the desired formula.

The enumerated properties enable us to compute the scalar product of linear combinations of vectors. For instance,

$$(a + 2b) \cdot (2a - 3b) = 2a \cdot a - 3a \cdot b + 4b \cdot a - 6b \cdot b =$$
$$= 2a^2 + a \cdot b - 6b^2$$

§ 3. Cartesian Coordinates in Space

9. **Cartesian Coordinates in Space.** Let us take a triad of vectors a, b and c drawn from a common origin at a point O and not lying in one and the same plane. We choose the point O as the origin of coordinates. The origin of a coordinate system given, the position of an arbitrary (variable) point M in space is completely specified by the vector $r = \overrightarrow{OM}$. This vector is called the **radius-vector** of the point M (see Fig. 156). As proved in Sec. 5, we can take the vectors a, b and c as a basis and represent the radius-vector in the form $r = \lambda a + \mu b + \nu c$. Thus, the position of the point M is characterized by the triple of numbers λ, μ and ν which are called the **affine coordinates** of the point M. In accordance with the end of Sec. 5 we can say that the affine coordinates of a point are the coordinates of its radius-vector. Therefore, every point in space, just as a point in a plane, has certain coordinates and, conversely, the

coordinates being given, we can always construct the corresponding point (but in space a point has three coordinates).

If vectors chosen as a basis of a coordinate system have unit lengths and are mutually perpendicular the coordinate system is called a **Cartesian system**. In such a case vectors forming a Cartesian
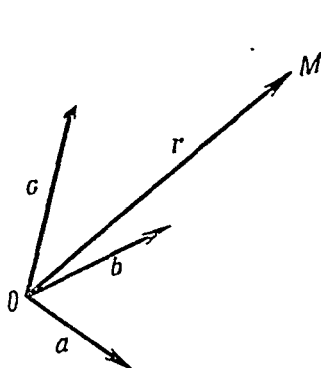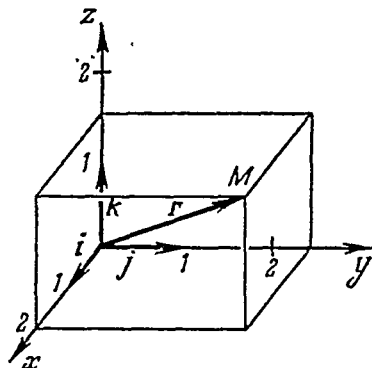
Fig. 156   Fig. 157

basis are usually designated by the letters **i**, **j** and **k**. Cartesian coordinates are usually denoted as x, y and z. Thus, according to Fig. 157, we have

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \qquad (7)$$

In Fig. 157 we see the point M with the coordinates $x = 2$, $y = 2.4$ and $z = 1.6$. The planes xOy, yOz and xOz (the coordinate planes) divide the space into eight parts (octants). The signs of the coordinates of a point show in which octant the point lies.

Similarly, any vector **a** can be represented (with respect to a coordinate system) in the form analogous to (7):

$$\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k} \qquad (8)$$

where a_x, a_y and a_z are the projections of the vector **a** on the corresponding coordinate axes. Taking into account that for any unit vector **e** we have, by formula (2), $\text{proj}_l \mathbf{e} = \cos(\widehat{\mathbf{e}, l})$, we deduce from formula (8) the relation

$$\mathbf{e} = \cos(\widehat{\mathbf{e}, x})\,\mathbf{i} + \cos(\widehat{\mathbf{e}, y})\,\mathbf{j} + \cos(\widehat{\mathbf{e}, z})\,\mathbf{k}$$

10. Some Simple Problems Concerning Cartesian Coordinates.
1. *Operations on vectors represented in terms of their projections on the axes of a Cartesian coordinate system* are performed according to the following simple rules. If [see formula (8)] we have

$$\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k} \quad \text{and} \quad \mathbf{b} = b_x\mathbf{i} + b_y\mathbf{j} + b_z\mathbf{k}$$

then (see Secs. 2, 4 and 8)

$$a + b = (a_x + b_x)\,i + (a_y + b_y)\,j + (a_z + b_z)\,k \qquad (10)$$

$$\lambda a = \lambda a_x i + \lambda a_y j + \lambda a_z k \qquad (11)$$

$$a \cdot b = a_x b_x + a_y b_y + a_z b_z \qquad (12)$$

$$a^2 = a \cdot a = a_x^2 + a_y^2 + a_z^2 \qquad (13)$$

To deduce the last two formulas which are of great importance we should notice that $i \cdot i = j \cdot j = k \cdot k = 1$ and $i \cdot j = j \cdot k = k \cdot i = 0$ since i, j and k are mutually perpendicular unit vectors. Formula (13) is nothing but the expression of Pythagoras' theorem in space: the square of the length of the diagonal of a rectangular parallelepiped equals the sum of the squares of the lengths of its sides.

For example, let it be required to determine the angle between the vectors $a = 3i - 2j + k$ and $b = -2i + j + 4k$. We have, by formula (3),

$$\cos(\widehat{a,\,b}) = \frac{a \cdot b}{ab} = \frac{a \cdot b}{\sqrt{a^2}\,\sqrt{b^2}} = \frac{3\,(-2) + (-2)\,1 + 1 \cdot 4}{\sqrt{3^2 + (-2)^2 + 1^2}\,\sqrt{(-2)^2 + 1^2 + 4^2}} =$$

$$= \frac{-4}{\sqrt{14 \cdot 21}} = -0.233$$

from which we obtain, within the accuracy guaranteed by a slide rule, $(\widehat{a,\,b}) = 103.5°$.

2. *The parallelism and perpendicularity conditions for vectors* given in the form of representation (9) can be derived as follows. The condition $a \parallel b$ is, by Sec. 5, equivalent to the relation of the form $b = \lambda a$ or, on the basis of formula (11), $b_x = \lambda a_x$, $b_y = \lambda a_y$, $b_z = \lambda a_z$. Now, eliminating λ, we get the desired condition

$$\frac{b_x}{a_x} = \frac{b_y}{a_y} = \frac{b_z}{a_z} \quad \text{(which is the condition of } a \parallel b\text{)}$$

Further, according to Sec. 8 (property 1), the condition $a \perp b$ is equivalent to the equality $a \cdot b = 0$ and, by formula (12), we deduce the condition

$$a_x b_x + a_y b_y + a_z b_z = 0 \quad \text{(which is the condition of } a \perp b\text{)}$$

3. The cosines of the angles which a vector forms with coordinate axes are called the **direction cosines** of the vector. If a vector a is given in the form of its representation (8) then we have, by formula (2), $a_x = \text{proj}_x\,a = a \cos(\widehat{a,\,x})$ etc., that is

$$\cos(\widehat{a,\,x}) = \frac{a_x}{a}, \quad \cos(\widehat{a,\,y}) = \frac{a_y}{a}, \quad \cos(\widehat{a,\,z}) = \frac{a_z}{a}$$

This implies

$$\cos^2{(\widehat{a,\ x})} + \cos^2{(\widehat{a,\ y})} + \cos^2{(\widehat{a,\ z})} = \frac{a_x^2}{a^2} + \frac{a_y^2}{a^2} + \frac{a_z^2}{a^2} = 1$$

The direction cosines of a vector completely determine its direction but give no information about its length.

The direction cosines of an axis are defined in like manner: they are the direction cosines of an arbitrarily taken vector parallel to the axis and having the same direction.
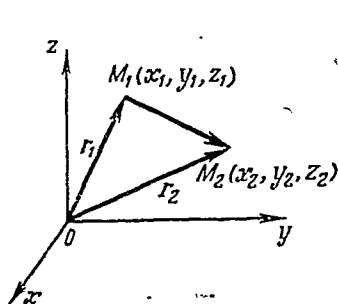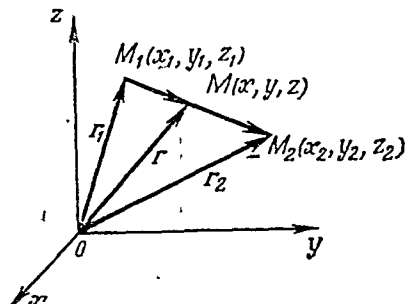
Fig. 158   Fig. 159

4. *The vector connecting two given points* M_1 $(x_1,\ y_1,\ z_1)$ and M_2 $(x_2,\ y_2,\ z_2)$ can be found as follows. According to Fig. 158 we have $\overrightarrow{OM_1} = \mathbf{r}_1 = x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k}$ and $\overrightarrow{OM_2} = \mathbf{r}_2 = x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k}$ which implies (see Sec. 3):

$$\overrightarrow{M_1M_2} = \mathbf{r}_2 - \mathbf{r}_1 = (x_2 - x_1)\,\mathbf{i} + (y_2 - y_1)\,\mathbf{j} + (z_2 - z_1)\,\mathbf{k}$$

5. *The distance between two points* M_1 $(x_1,\ y_1,\ z_1)$ and M_2 $(x_2,\ y_2,\ z_2)$ is equal to

$$M_1M_2 = \sqrt{(\overrightarrow{M_1M_2})^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \qquad (14)$$

which follows from the preceding argument. This formula is very much like the corresponding formula for a plane [see formula (II.1)].

6. *Dividing a segment in a given ratio.* Suppose we have

$$M_1(x_1,\ y_1,\ z_1), \quad M_2(x_2,\ y_2,\ z_2) \quad \text{and} \quad \frac{M_1M}{MM_2} = \lambda$$

(see Fig. 159). It is required to find the point M $(x,\ y,\ z)$. We have

$$\mathbf{r}_1 = x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k}; \quad \mathbf{r}_2 = x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k} \qquad (15)$$

The vectors $\overrightarrow{M_1M}$ and $\overrightarrow{MM_2}$ being parallel, we see that $\overrightarrow{M_1M} = \lambda\overrightarrow{MM_2}$. But $\overrightarrow{M_1M} = \mathbf{r} - \mathbf{r}_1$ and $\overrightarrow{MM_2} = \mathbf{r}_2 - \mathbf{r}$, that is $\mathbf{r} - \mathbf{r}_1 =$

$= \lambda (r_2 - r), \quad r - r_1 = \lambda r_2 - \lambda r$ and $r + \lambda r = r_1 + \lambda r_2.$ Hence,

$$r = \frac{r_1 + \lambda r_2}{1 + \lambda} \tag{16}$$

Equating the projections of both sides on the axes x, y and z [see formulas (7) and (15)] we finally deduce

$$x = \frac{x_1 + \lambda x_2}{1 + \lambda}, \quad y = \frac{y_1 + \lambda y_2}{1 + \lambda}, \quad z = \frac{z_1 + \lambda z_2}{1 + \lambda} \tag{17}$$

When passing from formula (16), which represents the solution of the problem in a vectorial form, to formula (17), we have projected
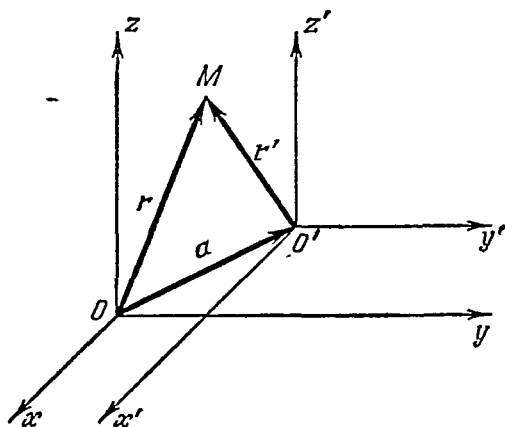
Fig. 160

the vectorial formula on the coordinate axes. Generally, it is clear that every vectorial equality of the form $a = b$ considered for vectors in space is equivalent to the three scalar equalities

$$a_x = b_x, \quad a_y = b_y, \quad a_z = b_z$$

which are obtained as a result of projecting the former equality on the coordinate axes.

7. *Translation of coordinate axes.* Let the coordinate axes x', y' and z' be obtained by means of a translation of the axes x, y and z, the displacement vector being a (see Fig. 160). Then we have the relation $r = a + r'$ for the radius-vectors of any point M. Projecting this relation on the coordinate axes we obtain the formula

$$x = x' + a_x, \quad y = y' + a_y, \quad z = z' + a_z$$

which describes the connection between the new coordinates and the old ones (see problem 3.I in Sec. II.2).

§ 4. Vector Product of Vectors

11. Orientation of Surface and Vector of an Area. A surface in space is called **oriented** if there is an indication as to which of its sides is regarded as outer and which as inner. As a rule, such an orientation can be performed in two ways (see Fig. 161). Even when we have a closed surface (e.g. a sphere) it is sometimes convenient
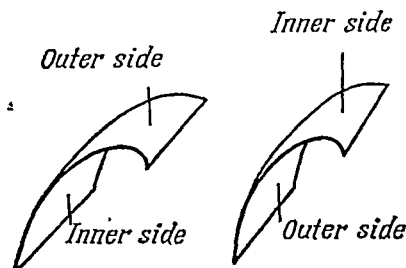
Fig. 161

to introduce an "unnatural" orientation, that is to regard the outer (in an ordinary, "everyday" sense) side of the surface as inner and vice versa.

An orientation of a non-closed surface can also be determined by pointing out the direction of describing its contour. Thus, we
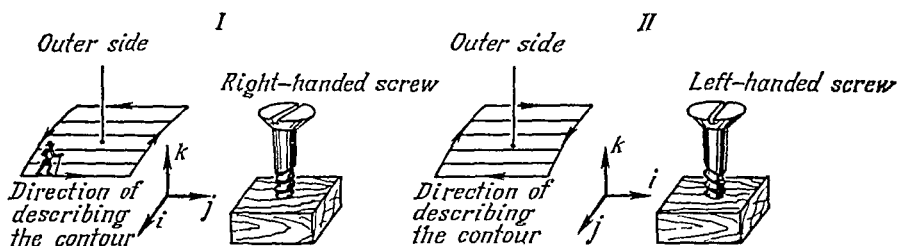
Fig. 162

have two methods of indicating an orientation of a non-closed surface. To establish the connection between them it is necessary to point out, in addition, whether we take the **right-hand screw rule** or the **left-hand screw rule** which are illustrated in Fig. 162. For instance, the right-hand screw rule can be stated in the following way: if we take a right-handed screw (which is usually used in engineering and everyday life) and rotate it in the direction of describing the contour it must move from the inner side of the surface to the outer side. Or, in other words, if we imagine that a man walks on

the outer side of the surface along its contour in the positive direction of describing the contour he must see "the precipice" on the right, and the surface itself should remain on the left.

When we consider an oriented part of a plane it sometimes turns out that only its area and its orientation in space are important whereas the specific form of the part (that is whether it is a circle or a rectangle etc.) does not matter. In such circumstances we can represent this part (S) of the plane by means of a vector which is perpendicular to (S) and is directed from its inner side to the outer side (see Fig. 163) whereas its absolute value is taken
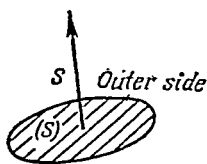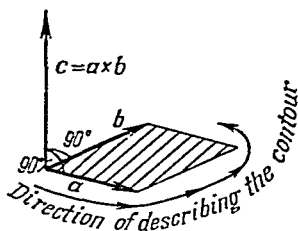
Fig. 163   Fig. 164

equal to the area of (S). We shall call such a vector the **vector of the area** (S) and denote it as S. Obviously, this vector completely defines the area and the orientation in space of such a surface element.

Let us discuss an example of using the vector of an area. Suppose we have a homogeneous gas flow, that is a flow in which the velocity **v** of the particles is the same at all points. Now imagine that we put an oriented plane surface element (S) into the flow, and it is required to determine the volume of the gas which passes through (S) during the unit interval of time from its inner side to the outer side. Since the volume of the gas passing in unit time fills a cylinder with base (S) and altitude $|v| \cos (\widehat{S, v})$ (think why it is so), the sought-for volume is equal to $|S| |v| \cos (\widehat{S, v}) = Sv$.

12. Vector Product. The vector product of two given vectors **a** and **b** is defined as the vector of the area of the parallelogram constructed on the vectors **a** and **b** (when the vectors are translated to a common origin) and oriented so that we should begin with the first vector (i.e. with **a**) while describing the contour of the parallelogram in the positive direction. The definition is illustrated in Fig. 164. We shall use the right-hand screw rule throughout our course unless the contrary is stated. We have also used the right-hand screw rule in Fig. 164.

Now let us introduce the notion of a **directed triad** of vectors which is important for our further purposes. Let three vectors **a**, **b** and **c** with a common origin be given. We shall regard the vectors as taken in a certain order, that is **a** is the first vector, **b** the second and **c** the third one. Let us also suppose that the vectors do not
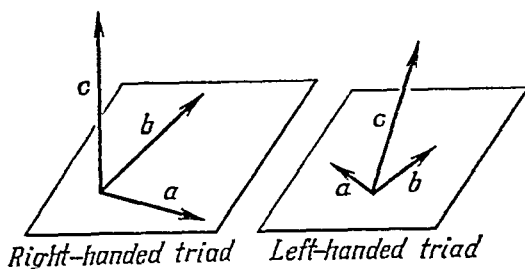
Fig. 165

lie in the same plane. Such a triad is said to be **right-handed** if the shortest rotation from the vector **a** to the vector **b** is seen to be in the counterclockwise direction when we contemplate the rotation from the terminus of the vector **c**. If the rotation is seen to be in the clockwise direction the triad is called **left-handed**. Right- and
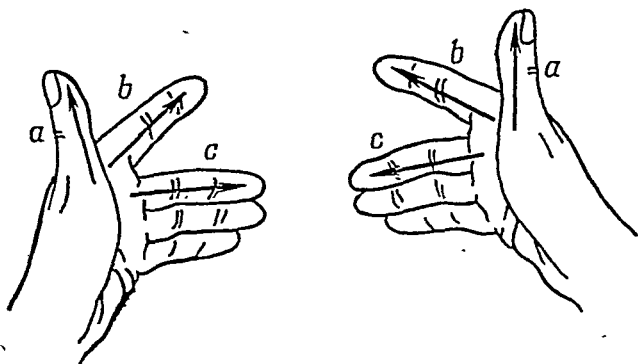
Fig. 166

left-handed triads are shown in Fig. 165. The origin of this terminology is illustrated in Fig. 166. Note that if we interchange the numbers of two vectors retaining the third one at its original place then the orientation of the triad will change. For instance, if a triad **a**, **b**, **c** is right-handed then the triad **a**, **c**, **b** is left-handed (check it!). The orientation of a triad does not change when we perform the so-called **circular permutation**, that is when the third vector is substituted for the second vector and the first vector is

substituted for the third one (or the other way round). For example, if a triad a, b, c is right-handed then the triad b, c, a is also right-handed.

The orientation of a Cartesian triad i, j, k must always correspond to the screw rule that we choose. Thus, i, j, k must form a right-handed triad in case the right-hand screw rule is chosen and they must form a left-handed triad if otherwise. In accordance with the rule we distinguish between the so-called **right-handed and left-handed Cartesian coordinate systems.** For instance, the Cartesian system in Fig. 157 is a right-handed one. Now we can formulate one more definition of a vector product which, as it can be seen in Fig. 164, is equivalent to the former definition.

The vector product of two vectors a and b is a vector c which is directed perpendicularly to either vector, has an absolute value equal to the area of the parallelogram constructed on the vectors a and b and forms a triad with the vectors a and b (the triad a, b, c) directed as the triad i, j, k (that is the triads a, b, c and i, j, k have the same orientation). Thus, the triad a, b, c is right-handed or left-handed in accordance with the Cartesian triad i, j, k. The vector product of the vectors a and b is denoted as $a \wedge b$ or [a, b].

13. Properties of Vector Product.

1. The absolute value of the vector product $a \times b$ is equal to $|a \times b| = ab \sin (\widehat{a, b})$. Indeed, this expression corresponds to the formula of calculating the area of a parallelogram.

The last formula together with formula (3) and property 2 in Sec. 8 implies the following consequence:

$$(a \times b)^2 + (a \cdot b)^2 = |a \times b|^2 + (a \cdot b)^2 =$$
$$= a^2 b^2 \sin^2 (\widehat{a, b}) + a^2 b^2 \cos^2 (\widehat{a, b}) = a^2 b^2$$

It should be understood that the first summand on the left-hand side is the scalar square of a vector (i.e. of $a \wedge b$) whereas the second one is the square of a scalar [of the scalar $(a \cdot b)$].

2. A vector product is equal to zero if and only if the vectors are parallel:

$$a \wedge b = 0 \text{ is equivalent to } a \parallel b$$

Indeed, if the vectors are parallel then the corresponding parallelogram degenerates into a line segment having the zero area and vice versa. In particular, we always have $a \times a = 0$.

3. The vector product is anticommutative:

$$b \times a = -(a \times b)$$

Really, changing the order of the factors does not affect the form of the parallelogram but its contour will be described in the opposite

direction and the vector of the surface element will therefore be replaced by the opposite one.

4. A scalar factor can be taken outside the vector product:

$$(\lambda a) \times b = a \times (\lambda b) = \lambda (a \times b)$$

since increasing the length of one of the sides of a parallelogram λ times results in increasing its area λ times. (If $\lambda < 0$ then the directions of the corresponding vectors are replaced by the opposite directions but the above rule remains true.)

5. Distributive law:

$$(a + b) \times c = a \times c + b \times c,$$
$$c \times (a + b) = c \times a + c \times b \qquad (18)$$

To prove the law let us translate the vectors a, b and c to a common origin O and draw a plane $(P) \perp c$ through O (see Fig. 167).
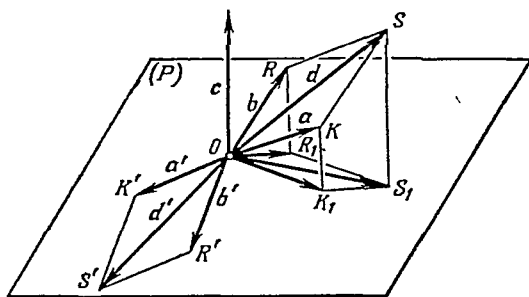
Fig. 167

$OK_1S_1R_1$ is the projection of the parallelogram $OKSR$; $OK'S'R$ is obtained by turning and stretching the parallelogram $OK_1S_1R_1$

Now we construct a parallelogram on the vectors a and b, its diagonal being d. Let us project the parallelogram on the plane (P). Next we turn the projected parallelogram about the axis c through 90° and simultaneously extend it c-fold (the whole construction is shown in Fig. 167). It is clear that

$$d = a + b, \quad d' = a' + b' \qquad (19)$$

where a′ and b′ are the sides of the third parallelogram and d′ is its diagonal.

For the sake of convenience the operations on the vector a performed in the above construction are depicted separately in Fig. 168. We see that $a' \perp OK$ and $a' \perp c$ [because $c \perp (P)$] and the vector a′ is therefore perpendicular to the plane KOM. Besides, $a' = cOK_1$, the last product being equal to the area of the parallelogram $OKLM$ constructed on the vectors a and c (why is it so?).

Consequently, by the definition of the vector product, $\mathbf{a}' = \mathbf{a} \times \mathbf{c}$. Similarly, $\mathbf{b}' = \mathbf{b} \times \mathbf{c}$ and $\mathbf{d}' = \mathbf{d} \times \mathbf{c}$. Now we deduce from formulas (19) the first formula (18): $(\mathbf{a} + \mathbf{b}) \times \mathbf{c} = \mathbf{d} \times \mathbf{c} = \mathbf{d}' = \mathbf{a}' + \mathbf{b}' = \mathbf{a} \times \mathbf{c} + \mathbf{b} \times \mathbf{c}$. Changing the order of the factors and simultaneously changing the signs we receive the second formula (18).

These properties enable us to remove brackets in expressions containing vector products but in doing this we should pay attention to the order of factors. Here we give an example:
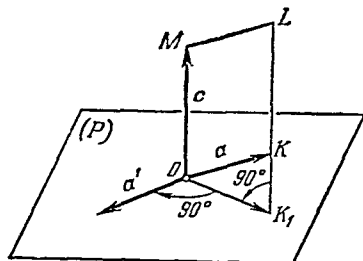
Fig. 168

$$(\mathbf{a} + 2\mathbf{b}) \times (2\mathbf{a} - 3\mathbf{b}) = 2\mathbf{a} \times \mathbf{a} - 3\mathbf{a} \times \mathbf{b} + 4\mathbf{b} \times \mathbf{a} - 6\mathbf{b} \times \mathbf{b} = -7\mathbf{a} \times \mathbf{b}$$

It should be noted that the associative law does not hold for the vector product: $(\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$ may be unequal to $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$. That is why expressions of the form $\mathbf{a} \times \mathbf{b} \times \mathbf{c}$ must not be used without brackets.

6. Expressing vector product in terms of Cartesian projections.

Let vectors $\mathbf{a}$ and $\mathbf{b}$ be given in the form of their representations (9). To express their vector product in terms of their Cartesian projections we shall utilize the following equalities:

$$\mathbf{i} \times \mathbf{j} = \mathbf{k}, \quad \mathbf{j} \times \mathbf{i} = -\mathbf{k}, \quad \mathbf{j} \times \mathbf{k} = \mathbf{i}, \quad \mathbf{k} \times \mathbf{j} = -\mathbf{i},$$
$$\mathbf{k} \times \mathbf{i} = \mathbf{j}, \quad \mathbf{i} \times \mathbf{k} = -\mathbf{j}$$

An essential fact is that these equalities do not depend on the particular choice of a Cartesian coordinate system and on the choice of the right-hand or left-hand screw rule. The verification of these equalities is left to the reader. Now we have

$$\begin{aligned}
\mathbf{a} \times \mathbf{b} &= (a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k}) \times (b_x\mathbf{i} + b_y\mathbf{j} + b_z\mathbf{k}) = \\
&= a_xb_y\mathbf{k} - a_xb_z\mathbf{j} - a_yb_x\mathbf{k} + a_yb_z\mathbf{i} + a_zb_x\mathbf{j} - a_zb_y\mathbf{i} = \\
&= \mathbf{i}(a_yb_z - a_zb_y) - \mathbf{j}(a_xb_z - a_zb_x) + \mathbf{k}(a_xb_y - a_yb_x)
\end{aligned} \quad (20)$$

The result thus obtained can be rewritten in the form of a determinant [see formula (VI.8)]:

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix}$$

This is a remarkable formula! In this form the formula for the vector product is easy to memorize.

Suppose it is necessary to evaluate the area S of a parallelogram constructed on the vectors $a = 3i - 2j + k$ and $b = -2i + j + 4k$. As we know, $S = |a \times b|$. Calculating we find

$$a \times b = \begin{vmatrix} i & j & k \\ 3 & -2 & 1 \\ -2 & 1 & 4 \end{vmatrix} = i(-8-1) - j(12+2) + k(3-4) =$$

$$= -9i - 14j - k$$

and thus

$$S = |a \times b| = \sqrt{9^2 + 14^2 + 1^2} = 16.7$$

The result is dimensionless since the vectors a and b were dimensionless. If we wanted to receive the "genuine area" we should put, for example, $a = 3i$ cm and the like.

Now we point out one more useful formula. Let us take two vectors $a = a_x i + a_y j$ and $b = b_x i + b_y j$ in the x, y-plane. Suppose it is necessary to evaluate the area S of the parallelogram constructed on a and b. Since

$$a \times b = \begin{vmatrix} i & j & k \\ a_x & a_y & 0 \\ b_x & b_y & 0 \end{vmatrix} = \begin{vmatrix} a_x & a_y \\ b_x & b_y \end{vmatrix} k$$

the geometrical meaning of the vector product implies (verify it!)

$$\begin{vmatrix} a_x & a_y \\ b_x & b_y \end{vmatrix} = \pm S, \quad \text{that is} \quad S = \left| \begin{vmatrix} a_x & a_y \\ b_x & b_y \end{vmatrix} \right| \quad (21)$$

We take $+$ or $-$ depending on whether the direction of the shortest rotation from a to b coincides with the direction of the rotation from i to j or not.

The notion of the **moment** of a vector a applied at a point M about a fixed point O is connected with the notion of a vector product. The moment $\text{mom}_O\, a$ is defined by the formula $\text{mom}_O\, a = \overrightarrow{OM} \times a = r \times a$. In physics we consider the moments of a force, of a velocity, of a momentum and so on. If we change the position of the origin of the vector a then, in general, its moment will also change. Let us now translate the vector a along its "line of action" to a point M'. Denote a in the new position as a'. Since $\overrightarrow{MM'} = \lambda a$ we have

$$\text{mom}_O\, a' = r' \times a = (r + \lambda a) \times a = r \times a = \text{mom}_O\, a$$

Thus, such a translation does not affect the moment and we may therefore regard the vector a entering in the definition of a moment as a sliding vector (see Sec. 1).

14. **Pseudovectors.** There are vectors which depend on whether we choose the right-hand or the left-hand screw rule so that when

one of the rules is replaced by the other the directions of these vectors are replaced by the opposite ones. Such vectors are called **pseudovectors (or axial vectors)** in contrast to true vectors whose direction is independent of the choice of a screw rule. For instance, the velocity vector of a translatory motion of a solid does not depend on the choice of a screw rule (this is implied by its physical meaning) and is therefore a true vector. On the contrary, the angular velocity vector ω of a rotary motion of a solid is a pseudovector. Indeed,
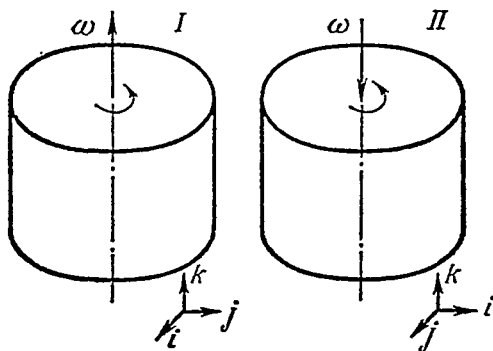
Fig. 169

such a vector lies in the axis of revolution and its absolute value is equal to the numerical value of the angular speed but its direction depends on the choice of a screw rule (see Fig. 169).

It is clear that by the definition of the vector product of two true vectors a vector product is a pseudovector because if we change the screw rule the former outer side of the parallelogram constructed on the vectors a and b becomes the inner side and vice versa. The same reasoning shows that the moment of a force (see the end of Sec. 13) is a pseudovector. Further, the vector product of a true vector by a pseudovector is a true vector whereas the vector product of two pseudovectors is also a pseudovector. It is also easy to verify that the linear velocity v of any point M of a rotating solid which is a true vector is connected with the pseudovector ω by the formula $\mathbf{v} = \omega \times \mathbf{r}$ where $\mathbf{r} = \overrightarrow{OM}$ and O is an arbitrarily chosen point lying on the axis of revolution.

We sometimes also distinguish between "true scalars" and **pseudoscalars**. A pseudoscalar is a scalar which is multiplied by -1 when the original screw rule is changed. For example, it is easy to verify that the scalar product of a true vector by a pseudovector is a pseudoscalar.

§ 5. Products of Three Vectors

15. Triple Scalar Product. Let three vectors a, b and c be given. The scalar quantity (a × b)·c is called the **triple (mixed) product** of the three vectors a, b and c. The geometrical significance of the triple product is seen in Fig. 170:

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{d} \cdot \mathbf{c} = d \operatorname{proj}_d \mathbf{c} = |\mathbf{a} \times \mathbf{b}| \operatorname{proj}_d \mathbf{c} = Sh = V$$

that is we have obtained the volume of a parallelepiped constructed on the vectors a, b and c. The vectors a, b and c in Fig. 170 form
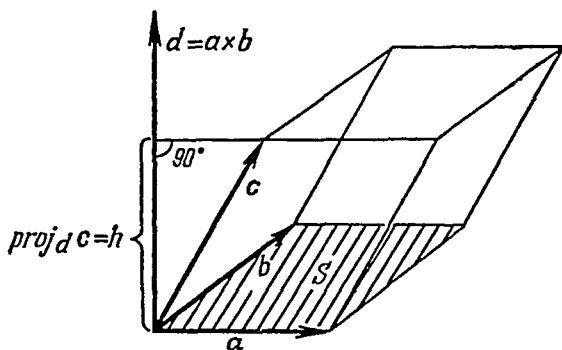
Fig. 170

a right-handed triad, and the volume is obtained with the sign $+$. If the triad were left-handed the angle between c and d would be obtuse. In this case $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = -V$. (We suppose here that our considerations are based on the right-hand screw rule as it was pointed out in Sec. 12.)

Let us indicate the following properties of a triple product.

1. A circular permutation of factors does not change their triple product since neither the parallelepiped (see Fig. 170) nor the orientation of the triad of the vectorial factors changes under such a transformation (see Sec. 12):

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = (\mathbf{b} \times \mathbf{c}) \cdot \mathbf{a} = (\mathbf{c} \times \mathbf{a}) \cdot \mathbf{b}$$

But if we interchange only two factors the sign of the triple product changes. For example, $(\mathbf{c} \times \mathbf{b}) \cdot \mathbf{a} = -(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$.

2. A triple product is equal to zero if and only if the three vectors are parallel to a plane. Indeed, such a parallelism means that the corresponding parallelepiped degenerates into a plane geometrical figure, that is it has a zero volume.

3. By formulas (20) and (12) we deduce the following expression of a triple product in terms of Cartesian projections:

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = [(a_y b_z - a_z b_y)\, \mathbf{i} - (a_x b_z - a_z b_x)\, \mathbf{j} +$$
$$+ (a_x b_y - a_y b_x)\, \mathbf{k}] \cdot (c_x \mathbf{i} + c_y \mathbf{j} + c_z \mathbf{k}) =$$
$$= (a_y b_z - a_z b_y)\, c_x - (a_x b_z - a_z b_x)\, c_y + (a_x b_y - a_y b_x)\, c_z$$

or, finally,

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix}$$

(to verify the formula expand the determinant in minors of the last row according to Sec. VI.3).

Let three vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ be given in the form of the expression in terms of their Cartesian projections. Then, by the above formula and property 2, we obtain the necessary and sufficient condition for these three vectors to be parallel to the same plane:

Fig. 171

$$\begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix} = 0 \qquad (22)$$
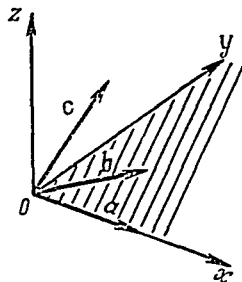
The triple product of three true vectors (see Sec. 14) is a scalar product of a pseudovector by a true vector, i.e. a pseudoscalar.

16. Triple Vector Product. The vector $(\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$ is called the triple vector product of three vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. It has no important geometrical meaning but is expressed by a formula which is of use for some applications. To deduce this formula let us choose the Cartesian coordinate axes in such a way that the x-axis is directed along the vector $\mathbf{a}$ and the y-axis lies in the plane of vectors $\mathbf{a}$ and $\mathbf{b}$ (see Fig. 171). Then the projections of the vector $\mathbf{a}$ on the y-axis and on the z-axis will be equal to zero, that is $\mathbf{a} = a_x \mathbf{i}$. Similarly, $\mathbf{b} = b_x \mathbf{i} + b_y \mathbf{j}$ and $\mathbf{c} = c_x \mathbf{i} + c_y \mathbf{j} + c_z \mathbf{k}$. From this we obtain

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_x & 0 & 0 \\ b_x & b_y & 0 \end{vmatrix} = a_x b_y \mathbf{k},$$

$$(\mathbf{a} \times \mathbf{b}) \times \mathbf{c} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & a_x b_y \\ c_x & c_y & c_z \end{vmatrix} = -\mathbf{i} a_x b_y c_y + \mathbf{j} a_x b_y c_x =$$
$$= a_x c_x (b_x \mathbf{i} + b_y \mathbf{j}) - (b_x c_x + b_y c_y)\, a_x \mathbf{i}$$

(check up these formulas!). Finally, using formula (12) we get

$$(a \times b) \times c = (a \cdot c) \, b - (b \cdot c) \, a$$

This final formula no longer contains any coordinate projections and therefore does not depend on the particular choice of the coordinate system.

The following formula is also sometimes of use:

$$a \times (b \times c) = -(b \times c) \times a = -[(b \cdot a) \, c - (c \cdot a) \, b] =$$
$$= (a \cdot c) \, b - (a \cdot b) \, c$$

In conclusion we note that vector algebra is a comparatively new branch of mathematics. It was created and developed in the second half of the 19th century in connection with problems of algebra, geometry, mechanics and physics.

§ 6. Linear Spaces

17. Concept of Linear Space. One of the characteristic features of vectors is the possibility to perform linear operations on them, that is the operation of addition and the operation of multiplication by numbers (see § 1). These operations can also be performed on some other objects, such as polynomials or arbitrary functions. Since these operations have similar properties in all cases, there is every reason to consider the general notion of a linear space which is understood as a set of some objects such that the linear operations can be performed on them within the set. Such a general, abstract, consideration yields a general view on linear operations which enables us to find some important properties in concrete problems.

Let (R) be a set (totality) of some objects. We shall call these objects elements (members) of (R). If a is one of the objects we also say that a belongs to (R). This fact is designated as $a \in (R)$. The fact that an element b does not belong to (R) is written as $b \bar{\in} (R)$. For instance, if (I) is the set of all integers then $3 \in (I)$, $-5 \in (I)$ but $\pi \bar{\in} (I)$.

Now we turn to the strict definition of the concept of a linear space. A set (R) is called a linear space if for any $x \in (R)$ and $y \in (R)$ the notion of the sum $x + y \in (R)$ is defined in a certain way, and if for any real number λ the product $\lambda x = x\lambda \in (R)$ is defined. For example, we can regard the set of all vectors for which the addition and the multiplication by numbers are performed in accordance with the rules of § 1 as a space (R). The set of all complex numbers for which the rules of addition and multiplication by real numbers are known from elementary mathematical courses can also be regarded as a linear space. Some other examples will be given in Sec. 18. Besides, the operations of addition and multiplication must have

some natural properties, namely those properties which were proved in § 1 for vectors and which should be introduced as the *axioms of a linear space* in the general case. The properties being essentially the same as those of vectors, the elements of a linear space are also often called vectors and are denoted like vectors.

Here we shall give the axioms of a linear space without discussing the question of their independence. The thing is that some of these axioms are the consequences of the rest but this is of no importance for the aims of our course. Thus, the sum must satisfy the following conditions:

1. Associativity: $(x + y) + z = x + (y + z)$ for any $x, y, z \in (R)$.
2. Commutativity: $x + y = y + x$ for any $x, y \in (R)$.
3. The existence of a zero element in (R): there must be an element [denoted as $0; 0 \in (R)$] which satisfies the condition $x + 0 = x$ for any $x \in (R)$.
4. The existence of the negative of $x \in (R)$: for any $x \in (R)$ there must be an element [denoted as $-x; -x \in (R)$] which satisfies the condition $(-x) + x = 0$.

It is easily verified that the zero element of a space must be unique and that there is only one negative for any element. We shall not discuss the general proof of these facts (in all the concrete examples which we shall consider here these properties are obvious).

The multiplication of an element by a number must satisfy the following requirements:

5. $\lambda (\mu x) = (\lambda \mu) x$.
6. $1x = x$.
7. $(-1) x = -x$.
8. $0x = 0$.
9. $\lambda 0 = 0$.

The operation of division by a number is introduced by the formula $\frac{x}{\lambda} = \frac{1}{\lambda} x$ $(\lambda \neq 0)$.

Finally, both linear operations are connected by the **distributive laws**:

10. $(\lambda + \mu) x = \lambda x + \mu x$ and 11. $\lambda (x + y) = \lambda x + \lambda y$.

All these properties make it possible to perform linear operations in linear spaces and transformations of linear combinations of elements (see Sec. 5) according to the usual arithmetical rules. Linear spaces and their properties are treated in detail in courses on linear algebra (see, for example, [16]).

It is sometimes necessary to consider sets of elements in which only one operation of addition satisfying axioms 1-4 is defined. Such a set is called an *Abelian group* after N. H. Abel (1802-1829), a Norwegian scientist and one of the most prominent mathematicians of the 19th century. Besides, it is sometimes possible to perform the multiplication not only by real numbers but also by any complex

numbers. A linear space of this kind is called *a linear space over the field of complex numbers* or, briefly, *a complex linear space*. The term "a number field" is applied to any set of numbers in which it is possible to perform the four fundamental operations of arithmetic with the natural exception of the division by zero. For example, all the real numbers or all the complex numbers form a field whereas the set of all integers does not form a number field (why is it so?).

18. Examples.

1. One of the simplest examples of a linear space is the set of all ordinary vectors with the linear operations described in § 1. The set of all vectors parallel to a certain plane is a linear space which is a linear subspace of the previous space. The set of all vectors parallel to a certain straight line is also a subspace of the space of all vectors. The zero vector itself is a linear space, from the formal point of view, since all the axioms 1-11 are fulfilled here.

In the general case a set (R_1) contained in a linear space (R) is called a linear subspace of (R) if (R_1) itself is a linear space with the operations which are originally defined for elements of (R). In other words, there must be $x + y \in (R_1)$ and $\lambda x \in (R_1)$ for any $x \in (R_1)$, $y \in (R_1)$ and an arbitrary real λ; if these two conditions hold the verification of axioms 1-11 is no longer needed since they are fulfilled in the whole space (R).

2. The set of all polynomials $P(x)$ of degree not higher than n where $n \geqslant 0$ is a given integer in a linear space. If n assumes successive values 0, 1, 2 etc. we get a sequence of linear spaces and each subsequent space contains all the preceding ones as linear subspaces. A still more general linear space is formed by the totality of all functions $f(x)$ defined over a fixed interval. Thus, each of these functions $f(x)$ can be regarded as a vector of the linear space of functions or, as we say, of a functional space. This approach is characteristic of modern mathematics.

3. The so-called n-dimensional real Cartesian space E_n where $n = 1, 2, 3, \ldots$ represents a very important example of a linear space and we shall consider this notion here.

We regard each ordered n-tuple $(a_1, a_2, \ldots, a_{n-1}, a_n)$ of real numbers $a_1, a_2, \ldots, a_{n-1}, a_n$ as an element of E_n. Take for example E_4. Then each element of the form (a_1, a_2, a_3, a_4) where a_1, a_2, a_3 and a_4 are arbitrary real numbers is a **point** or a **vector** of E_4 (ordinary geometrical vectors may be regarded as points of E_3 since we can regard them as radius-vectors of points; the vectors, i.e. the elements, of a general linear space are therefore also called points). The numbers a_1, a_2, a_3 and a_4 are called the coordinates of the point (of the vector) (a_1, a_2, a_3, a_4). For example, $(-3, 0, 1, 3)$ is one of such points. The point $(0, 0, 0, 0)$ is called the origin (of the coordinate system) in E_4. The whole space E_4 is the set, the totality, of all such points.

equation (23) we arrive at a linear dependence between x_1, x_2 and x_3. If $D = 0$ the right-hand sides of the first two relations (23) are proportional to each other and therefore even x_1 and x_2 are linearly dependent here, the same being true, of course, for x_1, x_2 and x_3.

Let us now consider as an example the linear space of all polynomials $P(x)$ of degree $\leqslant 2$ (see Sec. 18), that is of polynomials of the form $ax^2 + bx + c$. All the polynomials are linear combinations of the powers x^2, $x^1 = x$ and $x^0 = 1$, the powers themselves being linearly independent. Indeed, none of the powers is a linear combination of the rest. It follows that the space in question is three-dimensional. The above lemma implies that any set consisting of more than three such polynomials is linearly dependent. The elements 1, x and x^2 form a basis in this space.

Similarly, we conclude that the space of polynomials of degree $\leqslant n$ has the dimension $n + 1$. The space of polynomials of all degrees is infinite-dimensional.

The linear space E_n (see Sec. 18), as we could naturally expect, is n-dimensional in the sense of our definition. We shall illustrate this property by taking the space E_4 as an example. Let us introduce the following vectors e_1, e_2, e_3 and e_4:

$$e_1 \,(1,\ 0,\ 0,\ 0),\quad e_2\,(0,\ 1,\ 0,\ 0),\quad e_3\,(0,\ 0,\ 1,\ 0),$$
$$e_4\,(0,\ 0,\ 0,\ 1)$$

Obviously, the vectors are linearly independent. Besides, every vector $x\,(x_1,\ x_2,\ x_3,\ x_4)$ belonging to E_4 is represented as a linear combination of e_1, e_2, e_3 and e_4:

$$x\,(x_1,\ x_2,\ x_3,\ x_4) = x_1 e_1 + x_2 e_2 + x_3 e_3 + x_4 e_4$$

Hence, by the lemma, the linear space in question is four-dimensional. The vectors e_1, e_2, e_3 and e_4 form a basis in E_4.

There is an infinitude of bases in every finite-dimensional linear space. For instance, let us consider the space of polynomials of degree $\leqslant 2$. Choose three arbitrary values $x = x_1$, $x = x_2$, $x = x_3$ and denote the polynomial of the second degree which is equal to 1 at the point $x = x_k$ ($k = 1, 2, 3$) and equal to zero at other points x_l ($l = 1, 2, 3,\ l \neq k$) by $P_k(x)$. We can easily verify that $P_1(x) = \dfrac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)}$. The polynomials $P_2(x)$ and $P_3(x)$ are expressed similarly. The three polynomials $P_1(x)$, $P_2(x)$ and $P_3(x)$ are a basis in the space in question. The representation of an arbitrary polynomial of degree $\leqslant 2$ as a linear combination of $P_1(x)$, $P_2(x)$ and $P_3(x)$ is nothing but a special case of Lagrange's interpolation formula (V.23). Of course, we suppose here that $x_k \neq x_l$, $l, k = 1, 2, 3$, $k \neq l$.

We have already mentioned the notion of a linear subspace (R_1) (see Sec. 18) of a linear space (R). It is possible to prove that if

$(R_1) \neq (R)$ and if the space (R) is finite-dimensional the dimension of (R_1) is less than that of (R). The proof which we leave to the reader is rather simple; it is based on the definition of a dimension. By the way, it is convenient to use the following definition of the linear independence which is equivalent to the former definition when proving the above statement: vectors $\mathbf{a}$, $\mathbf{b}$, . . ., $\mathbf{d}$ are called linearly dependent if they satisfy a relation of the form $\alpha\mathbf{a} + + \beta\mathbf{b} + \ldots + \delta\mathbf{d} = 0$ where at least one of the numbers α, β, ..., δ is different from zero.

The important notion of a **hyperplane** is introduced with the aid of the notion of a linear subspace. Let us take a linear subspace of E_n and draw all the vectors of the subspace from a certain point of E_n. Then the set of points which are the termini of the vectors is a hyperplane. One-dimensional hyperplanes in E_n are called straight lines and two-dimensional hyperplanes in E_n are called planes. Besides, there are hyperplanes of dimensions 3, 4 etc. up to $n - 1$. We also introduce the formal notion of a "hyperplane of dimension 0" which is simply a separate point and the notion of a "hyperplane of dimension n" which is the whole space E_n. It is possible to construct "a stereometry" in E_n.

In conclusion we mention an important notion of an **isomorphism of linear spaces**. Two linear spaces (R) and (R') are called *isomorphic* (or, more precisely, *linearly isomorphic*) if it is possible to establish a one-to-one correspondence between the vectors of the spaces in such a way that the linear operations on the corresponding vectors are performed according to similar rules. A more comprehensive statement of this property is that if the vectors $\mathbf{x}$, $\mathbf{y} \in (R)$ correspond, respectively, to the vectors $\mathbf{x}'$, $\mathbf{y}' \in (R')$ then the vector $\mathbf{x} + \mathbf{y}$ must correspond to the vector $\mathbf{x}' + \mathbf{y}'$ and the vector $\lambda\mathbf{x}$ must correspond to the vector $\lambda\mathbf{x}'$, i.e.

$$(\mathbf{x} + \mathbf{y})' = \mathbf{x}' + \mathbf{y}', \quad (\lambda\mathbf{x})' = \lambda\mathbf{x}'$$

where the prime designates the transition from a vector of (R) to the corresponding vector of (R'). The term "one-to-one correspondence" means that only one vector from (R) corresponds to every vector from (R') and vice versa. Isomorphic linear spaces are indistinguishable from the point of view of linear operations since all the implications based on such operations in one of the spaces are true for all the isomorphic ones.

In particular, it follows that isomorphic linear spaces are of the same dimension. Conversely, all the finite-dimensional linear spaces of one and the same dimension are isomorphic to each other. Indeed, if $\mathbf{p}_1$, $\mathbf{p}_2$, . . ., $\mathbf{p}_n$ is a basis in (R) and $\mathbf{p}_1'$, $\mathbf{p}_2'$, . . ., $\mathbf{p}_n'$ is a basis in (R') it is easy to verify that the correspondence of the form

$$\alpha_1\mathbf{p}_1 + \alpha_2\mathbf{p}_2 + \ldots + \alpha_n\mathbf{p}_n \leftrightarrow \alpha_1\mathbf{p}_1' + \alpha_2\mathbf{p}_2' + \ldots + \alpha_n\mathbf{p}_n'$$

yields a desired isomorphism. Hence, with respect to linear operations, every finite-dimensional linear space is completely characterized by its dimension. In particular, we conclude that each finite-dimensional linear space of dimension n is isomorphic to the vector space E_n.

The theory of finite-dimensional linear spaces and its applications are studied in *linear algebra* whereas infinite-dimensional spaces are considered in *functional analysis*.

20. Concept of Euclidean Space. A linear space (R) is called a Euclidean space if the notion of a scalar product of any two vectors of (R) is introduced and if the following natural axioms of a scalar product are fulfilled:

12. The scalar product $x \cdot y = (x, y)$ of any two vectors $x, y \in (R)$ is a real number;

13. $(x, x) > 0$ for any $x \neq 0$;

14. $(x, y) = (y, x)$;

15. $(x + y, z) = (x, z) + (y, z)$;

16. $(\lambda x, y) = \lambda (x, y)$.

All these properties were proved for ordinary geometrical vectors in § 3. Thus, the set of all geometrical vectors is a Euclidean space. The set of all vectors parallel to a plane is also a Euclidean space. The same is true for the set of all vectors parallel to a straight line. It is clear that a linear subspace of a Euclidean space is always a Euclidean space.

The set of all vectors of the n-dimensional Cartesian space E_n represents an important example of a Euclidean space if we introduce the scalar product of any two vectors $x (x_1, x_2, \ldots, x_n)$ and $y (y_1, y_2, \ldots, y_n)$ by the formula

$$(x, y) = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \tag{24}$$

which is analogous to the well-known formula (12). It is easy to verify that all axioms 12-16 are fulfilled here.

It is possible to introduce the notion of a length or, as we often say, of a norm of any vector by means of the formula $|x| = = \sqrt{(x, x)}$. The norm of x is also denoted as $\| x \|$. Axiom 13 implies that the norm of any nonzero vector is positive. If the norm of a vector is equal to unity the vector is called a *normalized vector* (for ordinary geometrical vectors we use the term "a unit vector"). From axioms 14 and 16 it straightway follows that

$$| \lambda x | = \sqrt{(\lambda x, \lambda x)} = \sqrt{\lambda^2 (x, x)} = |\lambda| \sqrt{(x, x)} = |\lambda| |x|$$

This implies, in particular, that $|0| = 0$ and that each vector $x \neq 0$ can be normalized, that is we obtain a normalized vector by dividing x by $|x|$.

Let us now deduce an estimate for a scalar product of two arbitrary vectors x, y $\in (R)$. In order to do this we note that we have

$$0 \leqslant |x + \lambda y|^2 = (x + \lambda y, x + \lambda y) =$$
$$= (y, y) \lambda^2 + 2 (x, y) \lambda + (x, x) \qquad (25)$$

for any λ. If we regard the right-hand side of (25) as a quadratic trinomial in λ the retention of its sign implies that the discriminant of the trinomial is non-positive, i.e.

$$(x, y)^2 - (x, x) (y, y) \leqslant 0$$

(why is it so?). From this we receive

$$|(x, y)| = \sqrt{(x, y)^2} \leqslant \sqrt{(x, x) (y, y)} = |x| |y| \qquad (26)$$

This important inequality was established in 1821 for E_n by Cauchy (Cauchy in fact used a terminology different from that of the theory of linear spaces). For some other cases the inequality was deduced in 1859 by V. Ya. Bunyakovsky (1804-1889), a prominent Russian mathematician, and by the German mathematician H. Schwarz (1843-1921) in 1884.

By inequality (26), we obtain, in particular,

$$|x + y|^2 = (x + y, x + y) = (x, x) + 2 (x, y) + (y, y) \leqslant$$
$$\leqslant |x|^2 + 2 |x| |y| + |y|^2$$

which implies

$$|x + y| \leqslant |x| + |y|$$

The last inequality is called the **triangle inequality** (think why it is called so).

A Euclidean space over the complex number field (see Sec. 17) or, as we briefly call it, a *complex Euclidean space* is also considered in mathematics. In this case a scalar product of two vectors can be a complex number and axiom 14 is therefore replaced by a new axiom of the form $(x, y) = (y, x)^*$ where the asterisk indicates the conjugate complex number. In particular, this implies that $(x, \lambda y) = (\lambda y, x)^* = [\lambda (y, x)]^* = \lambda^* (x, y)$. We can easily verify that all other assertions of this section remain true for complex spaces. The n-dimensional complex Cartesian space Z_n with the vectors of the form x $(x_1, x_2, \ldots, x_n)$ where $x_1, \ldots, x_n$ are arbitrary complex numbers is an important example of a complex Euclidean space if we introduce a scalar product by means of the formula $(x, y) = x_1 y_1^* + x_2 y_2^* + \ldots + x_n y_n^*$ which substitutes for formula (24). We leave to the reader the verification of all the axioms and, in particular, axiom 13 for this case.

21. Orthogonality. The notion of an angle between any two vectors is introduced in a natural way in a Euclidean space (R) by the

formula

$$\cos (\widehat{\mathbf{x}, \mathbf{y}}) = \frac{(\mathbf{x}, \mathbf{y})}{|\mathbf{x}| |\mathbf{y}|} \tag{27}$$

On the basis of Sec. 7 we see that the definition yields a usual angle in the case of ordinary geometrical vectors. It is necessary to note that, by inequality (26), the right-hand side of equality (27) does not exceed unity in its absolute value and therefore definition (27) always makes sense. In case vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent, i.e. $\mathbf{y} = \lambda \mathbf{x}$, formula (27) implies $(\widehat{\mathbf{x}, \mathbf{y}}) = 0°$ if $\lambda > 0$ and $(\widehat{\mathbf{x}, \mathbf{y}}) = 180°$ if $\lambda < 0$. Conversely, if $\cos (\widehat{\mathbf{x}, \mathbf{y}}) = \pm 1$ then inequality (27) shows that the vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent, but we leave the proof to the reader.

In the case

$$(\mathbf{x}, \mathbf{y}) = 0 \tag{28}$$

we have an important situation when the vectors $\mathbf{x}$ and $\mathbf{y}$ form an angle of 90° or, as it is often said, the vectors are **orthogonal**. If at least one of the vectors $\mathbf{x}$ or $\mathbf{y}$ is equal to 0 relation (28) necessarily holds and therefore the zero vector is regarded as orthogonal to any given vector although we can attribute any value to the angle which the zero vector forms with a given vector.

A system of vectors $\mathbf{a}$, $\mathbf{b}$, ..., $\mathbf{d}$ is called orthogonal if all the vectors are mutually pairwise orthogonal and none of them is equal to the zero vector. Such a system is always linearly independent. In fact, if we take the scalar product of an equality of the form $\mathbf{d} = \alpha \mathbf{a} + \beta \mathbf{b} + \ldots$ by $\mathbf{a}$ we receive $0 = \alpha (\mathbf{a}, \mathbf{a})$ which implies $\alpha = 0$. Similarly, we conclude that all other coefficients on the right-hand side of the last linear combination are equal to zero which is impossible.

It is most convenient to use an **orthogonal basis** in a finite-dimensional space, that is a basis which is simultaneously an orthogonal system. In fact, if $\mathbf{p}_1$, $\mathbf{p}_2$, ..., $\mathbf{p}_n$ is such a basis then any vector $\mathbf{x}$ is represented in the form $\mathbf{x} = \alpha_1 \mathbf{p}_1 + \alpha_2 \mathbf{p}_2 + \ldots + \alpha_n \mathbf{p}_n$, and to find the coefficients of the representation it is sufficient to multiply (in the sense of a scalar product) both sides by the vectors $\mathbf{p}_k$ ($k = 1, 2, \ldots, n$). By the orthogonality, this yields $(\mathbf{x}, \mathbf{p}_k) = \alpha_k (\mathbf{p}_k, \mathbf{p}_k)$ which implies

$$\alpha_k = \frac{(\mathbf{x}, \mathbf{p}_k)}{(\mathbf{p}_k, \mathbf{p}_k)} \quad (k = 1, 2, \ldots, n) \tag{29}$$

If we had a non-orthogonal basis the above method would give a system of n equations of the first degree in n unknowns α_k. Formula (29) becomes especially simple when the vectors which form an orthogonal basis are normalized (see Sec. 20). Indeed, then the

right-hand sides of (29) have denominators equal to unity. An orthogonal basis of normalized vectors is called a **Euclidean** basis (compare with Sec. 9) or an **orthonormal** basis.

An orthogonal basis can be constructed in any finite-dimensional Euclidean space. Actually, let, for example, (R) be four-dimensional and let q_1, q_2, q_3 and q_4 be a basis in (R). We put $p_1 = q_1$ and $p_2 = q_2 + \alpha p_1$ and try to choose α so that (p_1, p_2) should be equal to zero: $(p_2, p_1) = 0$. This yields $(q_2 + \alpha p_1, p_1) = 0$, i.e. $\alpha = -(q_2, p_1)(p_1, p_1)^{-1}$. Then we put $p_3 = q_3 + \beta_1 p_1 + \beta_2 p_2$ and choose β_1 and β_2 in such a way that $(p_3, p_1) = 0$ and $(p_3, p_2) = 0$. This implies (check it!) $\beta_1 = -(q_3, p_1)(p_1, p_1)^{-1}$ and $\beta_2 = -(q_3, p_2)(p_2, p_2)^{-1}$. Finally, putting $p_4 = q_4 + \gamma_1 p_1 + \gamma_2 p_2 + \gamma_3 p_3$ we determine the coefficients γ_k $(k = 1, 2, 3)$ in such a way that $(p_4, p_1) = 0$, $(p_4, p_2) = 0$ and $(p_4, p_3) = 0$. We leave the determination of γ_k $(k = 1, 2, 3)$ to the reader. It is easy to show that by the linear independence of the vectors q_k $(k = 1, 2, 3, 4)$ all the denominators entering in this procedure are different from zero and this orthogonalization process can therefore be completed. The vectors p_1, p_2, p_3 and p_4 thus obtained form an orthogonal basis in (R).

It follows that all the Euclidean spaces of the same dimension are isomorphic to each other. Let us discuss here this important corollary. Two given Euclidean spaces are called **isomorphic** if it is possible to establish a linear isomorphism (see Sec. 19) between them which preserves the scalar product, that is $(x, y)_R = (x', y')_{R'}$ (the subscripts R and R' indicate here the spaces in which the scalar products are taken). Two isomorphic Euclidean spaces are indistinguishable from the point of view of the theory of Euclidean spaces. It is obvious that two isomorphic Euclidean spaces are of the same dimension. But it is easy to show that, conversely, two Euclidean spaces of the same dimension are isomorphic. To prove this we choose an orthonormal basis in each of the spaces (by the way, if we normalize orthogonal vectors their orthogonality is preserved). Let these bases be p_1, p_2, ..., p_n and p_1', p_2', ..., p_n', respectively. Now we establish an isomorphism by using the procedure described in Sec. 19:

$$\alpha_1 p_1 + \alpha_2 p_2 + \ldots + \alpha_n p_n \leftrightarrow \alpha_1 p_1' + \alpha_2 p_2' + \ldots + \alpha_n p_n'$$

Actually, if $x = \alpha_1 p_1 + \ldots + \alpha_n p_n$ and $y = \beta_1 p_1 + \ldots + \beta_n p_n$ then

$$(x, y)_R = \alpha_1 \beta_1 + \ldots + \alpha_n \beta_n = (x', y')_{R'}$$

(why is it so?). In particular, we conclude that every finite-dimensional Euclidean space of dimension n is isomorphic to the vector space E_n. The same is also true for complex spaces.

§ 7. *Vector Functions of Scalar Argument. Curvature*

22. Vector Variables. Let us return to usual geometrical vectors
Vector variables are closely related to scalar variables. Indeed
if, for example, we introduce a Cartesian coordinate system and
write

$$\mathbf{u} = u_x\mathbf{i} + u_y\mathbf{j} + u_z\mathbf{k} \qquad (30$$

then it is seen that every change of the vector quantity **u** reduce
to certain changes of the scalar quantities u_x, u_y and u_z. We shal
give only a few simple remarks.

A vector quantity is infinitesimal (we denote this as $\mathbf{u} \to 0$
if its absolute value $|u|$ is an infinitesimal variable, i.e. $u \to 0$
But the direction of the vector **u** may change in an arbitrar
way in such a process and may not have a limit.

The vectorial limiting relationship $\mathbf{u} \to \mathbf{a}$ (where **a** is a constan
vector) is equivalent to the three scalar relationships $u_x \to a_x$
$u_y \to a_y$ and $u_z \to a_z$. In addition, if $\mathbf{a} \neq 0$ then **u** tends to a no
only in its absolute value but also in its direction. The theorem
on the limits of a sum, of a product and the like (see Sec. III.5
remain true here and their proofs hold too. But, of course, the pro
perties of limits which are connected with inequalities do not appl
to vectors since, as it was agreed before, we do not consider th
notion of an inequality for vectors in our course.

23. Vector Functions of Scalar Argument. We say that there i
a *vector function of a scalar argument* $\mathbf{u} = \mathbf{f}(t)$ if to each valu
of the scalar variable t there corresponds, by a certain law, a valu
of the vector quantity **u**.

According to the beginning of Sec. 22, the determination of on
vector function is equivalent to the determination of three scala
functions (understood in the ordinary sense) because $u_x = f_x(t)$
$u_y = f_y(t)$ and $u_z = f_z(t)$.

The concept of continuity of a vector function of a scalar argu
ment is introduced in the same way as for scalar functions. Further

$$\mathbf{u}' = \mathbf{f}'(t) = \lim_{\Delta t \to 0} \frac{\Delta \mathbf{u}}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbf{f}(t + \Delta t) - \mathbf{f}(t)}{\Delta t}$$

All the basic properties of a derivative (see Sec. IV.4) are transferred
(together with their proofs) to the case of vector functions. But i
is necessary to stress that while applying these rules to differentiating
a vector product we must pay attention to the order of factors
$(\mathbf{u} \times \mathbf{v})' = \mathbf{u}' \times \mathbf{v} + \mathbf{u} \times \mathbf{v}'$. The concept of Taylor's series is
also extended to vector functions (see Eq. IV.52).

The following property is sometimes of use: if $|\mathbf{u}(t)| = \text{const}$
then $\mathbf{u}' \perp \mathbf{u}$. Indeed, the condition can be written in the form

$\mathbf{u} \cdot \mathbf{u} = u^2 = \text{const.}$ Now differentiating we obtain $\mathbf{u'u} + \mathbf{uu'} = 0$ and $2\mathbf{u'u} = 0$ which implies $\mathbf{u'} \perp \mathbf{u}$.

To illustrate the geometrical meaning of a vector function of a scalar argument let us draw the vector $\mathbf{u}$ from a fixed point O in space. Then it is natural to regard $\mathbf{u}$ as a radius-vector (see Sec. 9) and denote it by $\mathbf{r}$, i.e.

$$\mathbf{r} = \mathbf{f}\,(t) \qquad \qquad \cdot \ (31)$$

As t varies the terminus of the vector $\mathbf{r}$ describes a curve (L) in space (see Fig. 172). We can therefore regard (31) as a *vector-parametric equation of the curve* (L). But if we introduce Cartesian axes
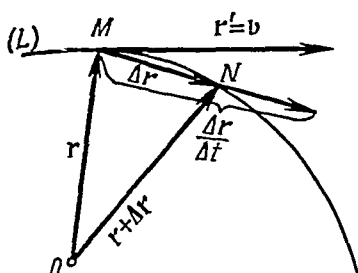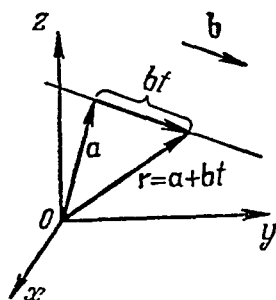
Fig. 172   Fig. 173

it is easy to pass from (31) to scalar-parametric equations of the form

$$x = \varphi\,(t), \quad y = \psi\,(t), \quad z = \chi\,(t) \qquad (32)$$

which determine the same curve. The right-hand sides of (32) are the result of projecting the function $\mathbf{f}\,(t)$ on the coordinate axes. [Compare this with parametric equations (II.10) of a plane curve.] If a space curve is originally defined by equations of form (32) then we pass to form (31) according to the formula $\mathbf{r} = \varphi\,(t)\,\mathbf{i} + \psi\,(t)\,\mathbf{j} + \chi\,(t)\,\mathbf{k}$. It is convenient to interpret the argument t as time. Then the curve (L) may be regarded as a trajectory of a moving point. This curve is also called the **hodograph** of the vector function $\mathbf{u}$.

For example, as it is shown in Fig. 173, the equation $\mathbf{r} = \mathbf{a} + \mathbf{b}t$ where $\mathbf{a}$ and $\mathbf{b}$ are some constant vectors determines a straight line and describes a uniform rectilinear motion along this line with the velocity $\mathbf{b}$. Projecting this equation on the coordinate axes we arrive at the *parametric equations of the straight line*:

$$x = a_x + b_x t, \quad y = a_y + b_y t, \quad z = a_z + b_z t \qquad (33)$$

As another example, let us take the *equation of a screw line* (*circular helix*). This curve can be obtained as a superposition of two motions of a point: the uniform motion along an axis and the uniform

rotation about the same axis (see Fig. 174). We choose the z-axis as the axis of revolution and denote the speed of the rectilinear motion by v and the angular speed by ω. Then we have

$$x = R \cos \omega t, \quad y = R \sin \omega t, \quad z = vt$$

or, in the form of a vector equation,

$$\mathbf{r} = R \ (\cos \omega t \ \mathbf{i} + \sin \omega t \ \mathbf{j}) + vt \ \mathbf{k}$$

If the variable t receives an increment Δt the point M on the curve (L) passes to the position N (see Fig. 172). Hence, $\Delta \mathbf{r} = \overrightarrow{MN}$.

The ratio $\frac{\Delta \mathbf{r}}{\Delta t}$ (which is a vector representing the **average velocity**) also lies on the straight line MN because Δt is a scalar. When $\Delta t \to 0$ we have $\frac{\Delta \mathbf{r}}{\Delta t} \to \mathbf{r}' = \mathbf{v}$ (where $\mathbf{v}$ is the **instantaneous velocity**). Let us prolong the secant MN and watch the change of the position of this straight line as $\Delta t \to 0$, i.e. as $N \to M$. The secant rotates about the point M in this process and tends to the position of the tangent to the trajectory at the point M. Hence we see that the instanta-neous velocity vector $\mathbf{v} = \frac{d\mathbf{r}}{dt}$ is directed along the tangent line to the trajectory, this fact being well known in mechanics. The vector-parametric equation of a tangent line drawn for a certain value $t = t_0$ has the form (see Fig. 173)
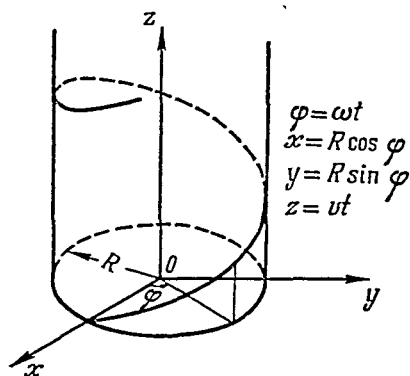
$$\mathbf{r} = \mathbf{r}_0 + \mathbf{v}_0 \ (t - t_0)$$

where $\mathbf{r}_0 = \mathbf{f} \ (t_0)$ and $\mathbf{v}_0 = \mathbf{f}' \ (t_0)$. We can rewrite this equation as follows: $\Delta \mathbf{r} = d\mathbf{f}$. But if we take the equation $\mathbf{r} = \mathbf{f} \ (t)$ of the curve (L) then $\Delta \mathbf{r} = \Delta \mathbf{f}$. Thus we see that the replacement of $\Delta \mathbf{f}$ by $d\mathbf{f}$ is equivalent to the replacement of the motion along the curve (L) by the uniform motion along the tangent line with the velocity equal to the instantaneous velocity at the given moment of time, that is by a motion which would appear if all the forces stopped acting on the moving point at this moment.

According to Sec. III.9 (see example 1) we have, as $\Delta s \to 0$,

$$\left| \frac{\Delta s}{\Delta \mathbf{r}} \right| \to 1, \quad \left| \frac{\Delta \mathbf{r}}{\Delta s} \right| \to 1, \quad \left| \frac{d\mathbf{r}}{ds} \right| = \lim \left| \frac{\Delta \mathbf{r}}{\Delta s} \right| = 1, \quad |d\mathbf{r}| = |ds| \qquad (34)$$

$$\varphi = \omega t$$
$$x = R \cos \varphi$$
$$y = R \sin \varphi$$
$$z = vt$$

Fig. 174

Here we take the absolute values of Δs and ds since these quantities can be both positive and negative. Let us choose now the arc length s as a parameter varying along the curve (L) and reckon it from a point M_0 of the curve, that is take the equation of the curve in the form $\mathbf{r} = \mathbf{r}\,(s)$. Then we see that the derivative $\frac{d\mathbf{r}}{ds}$ is a unit vector in the direction of the tangent to (L), that is the unit tangent vector (see Sec. 7). This vector is usually denoted by the letter $\boldsymbol{\tau}$. i.e. $\frac{d\mathbf{r}}{ds} = \boldsymbol{\tau}$. Consequently, we have

$$\frac{d\mathbf{r}}{dt} = \mathbf{v} = \frac{d\mathbf{r}}{ds}\frac{ds}{dt} = v\boldsymbol{\tau}, \qquad \boldsymbol{\tau} = \frac{\frac{d\mathbf{r}}{dt}}{\frac{ds}{dt}} = \frac{\dot{\mathbf{r}}}{\dot{s}} \tag{35}$$

We remark in conclusion that formula (34) implies the following expression for the differential of the arc length in Cartesian coordinates:

$$ds = \pm\,|\,d\mathbf{r}\,| = \pm\,|\,d\,(x\mathbf{i} + y\mathbf{j} + z\mathbf{k})\,| = \pm\,|\,dx\mathbf{i} + dy\mathbf{j} + dz\mathbf{k}\,| =$$
$$= \pm\,\sqrt{dx^2 + dy^2 + dz^2}$$

24. Some Notions Related to the Second Derivative. Since $|\,\boldsymbol{\tau}\,(s)\,| = 1 = \text{const}$ we have $\frac{d\boldsymbol{\tau}}{ds} \perp \boldsymbol{\tau}$ (see the beginning of Sec. 23). The straight line pp (see Fig. 175) drawn through a moving point
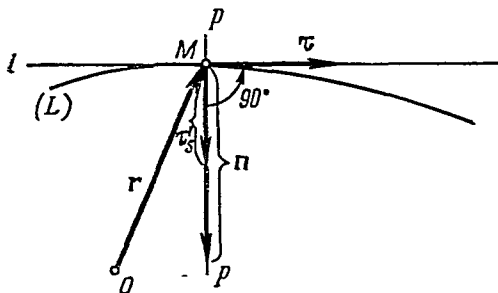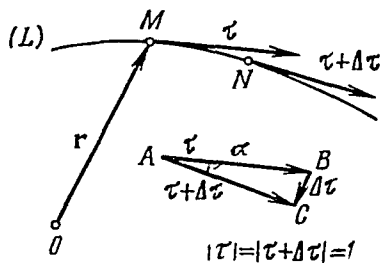
Fig. 175   Fig. 176

M of the trajectory (L) and parallel to $\frac{d\boldsymbol{\tau}}{ds}$ is therefore a normal to (L) (a perpendicular to the tangent ll at the point M). There exists an infinitude of normals drawn to a curve in space at each of its points. At each point these normals form a plane which is called the **normal plane**. To distinguish the normal in the direction of the vector $\frac{d\boldsymbol{\tau}}{ds}$ from other normals at a given point M we call this normal

the **principal normal** of the curve (L) at the point M. The length (the absolute value) of the vector $\frac{d\tau}{ds}$ is called the **curvature of the curve** (L) at the point M and is denoted by the letter k, i.e. $\left|\frac{d\tau}{ds}\right| = k$ and $\frac{d\tau}{ds} = k\mathbf{n}$ where $\mathbf{n}$ is the unit vector in the direction of the principal normal.

The geometrical significance of the curvature is shown in Fig. 176:

$$k = \left|\frac{d\tau}{ds}\right| = \lim\left|\frac{\Delta\tau}{\Delta s}\right| = \lim\frac{BC}{\Delta s} = \lim\frac{2\sin\frac{\alpha}{2}}{\Delta s} = \lim\frac{\alpha}{\Delta s}$$

(in the last passage to the limit we have used property 4 from Sec. III.8).

Thus, the curvature is the speed (the angular speed related to the unit of the distance passed over) with which the tangent to the curve rotates. Incidentally, we see that the vectors τ'_s and $\mathbf{n}$ go in the direction of the concavity of the curve.

Differentiating the first formula (35) with respect to t we obtain

$$\frac{d^2\mathbf{r}}{dt^2} = \frac{dv}{dt}\,\tau + v\,\frac{d\tau}{dt} = \frac{dv}{dt}\,\tau + v\,\frac{d\tau}{ds}\,\frac{ds}{dt} = \frac{dv}{dt}\,\tau + v^2 k\mathbf{n}$$

This formula is widely applied in mechanics since if t is the time of motion the formula shows that the acceleration vector $\frac{d^2\mathbf{r}}{dt^2}$ can be resolved into the components $\frac{dv}{dt}\,\tau$ and $v^2 k\mathbf{n}$. $\frac{dv}{dt}\,\tau$ is the *tangential component* since it has the direction of the tangent and $v^2 k\mathbf{n}$ is the *normal component* (directed along the principal normal).

Thus, the vector $\frac{d^2\mathbf{r}}{dt^2}$ drawn from a point M must necessarily lie in the plane passing through the tangent and the principal normal drawn from this point. This plane is called the **osculating plane** of the curve (L) at the point M. Applying Taylor's formula (IV.50) to $\mathbf{f}\,(t_0 + \Delta t)$ we conclude that the curve (L) may be regarded in the vicinity of its point as lying in its osculating plane with an accuracy up to infinitesimals of the second order (relative to Δt) inclusive. (It can similarly be shown that by formula (IV.49) the curve (L) coincides with its tangent to within infinitesimals of the second order relative to Δt as $\Delta t \to 0$.)

Hence, the osculating plane of the curve (L) at the point M may be regarded as a plane passing through three points of the curve (L) lying infinitely close to the point M (just as we regard the tangent line as a line passing through two points of a curve which are infinitely close to each other).

25. Osculating Circle. Let a curve (L) in a plane be represented parametrically (see Sec. II.6): $x = x\,(t)$ and $y = y\,(t)$. Then, as

it is seen in Fig. 177, the curvature k of the curve at any moving

point is equal to $\left| \dfrac{d\varphi}{ds} \right| = \left| \dfrac{d \, \text{arc} \tan \dfrac{\dot{y}}{\dot{x}}}{ds} \right|$ where the dot denotes the differentiation with respect to the parameter. Taking the expression of ds obtained in the end of Sec. 23 and performing some trans-
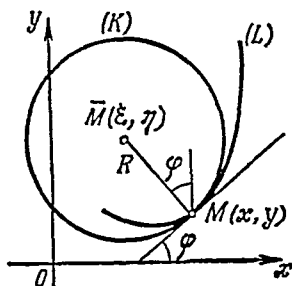
Fig. 177

$\tan \varphi = \dfrac{dy}{dx} = \dfrac{\dot{y}}{\dot{x}}$ (see Sec. IV.9)
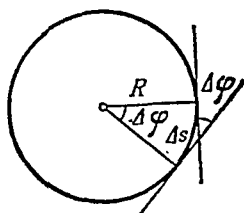
Fig. 178

$\Delta s = R \Delta \varphi$

formations we receive

$$k = \left| \frac{1}{1 + \left(\dfrac{\dot{y}}{\dot{x}} \right)^2} \frac{\ddot{y}\,\dot{x} - \dot{y}\,\ddot{x}}{\dot{x}^2} \, dt \right| : \sqrt{dx^2 + dy^2} =$$

$$= \left| \frac{1}{\dot{x}^2 + \dot{y}^2} (\ddot{y}\,\dot{x} - \ddot{y}\,\dot{x}) \, dt \right| : \sqrt{\dot{x}^2 + \dot{y}^2} \, dt = \frac{|\ddot{y}\,\dot{x} - \dot{y}\,\ddot{x}|}{(\dot{x}^2 + \dot{y}^2)^{3/2}} \qquad (36)$$

In particular, if the equation of a curve is represented in the form $y = f(x)$, that is the argument x itself is regarded as a parameter, then $\dot{y} = y'$, $\ddot{y} = y''$, $\dot{x} = 1$, $\ddot{x} = 0$ and

$$k = \frac{|y''|}{(1 + y'^2)^{3/2}} \qquad (37)$$

Fig. 178 shows that for a circle

$$k = \left| \frac{d\varphi}{ds} \right| = \lim \left| \frac{\Delta\varphi}{\Delta s} \right| = \lim \frac{\Delta\varphi}{R\Delta\varphi} = \frac{1}{R} \qquad (38)$$

which means that the curvature of a circle is constant and inverse to its radius. The only plane curves with constant curvature are circles and straight lines; the curvature of a straight line is equal to zero.

Let us take an arbitrary point M on a curve (L). The circle (K) passing through M and having the same tangent, the same curvature and the same direction of convexity as (L) is called the **osculating circle** (the "circle of curvature") of the curve (L) at the point M (see Fig. 177). The radius and the centre of this circle are called the **radius of curvature** and the **centre of curvature**, respectively. According to formulas (36)-(38) we have

$$R = \frac{1}{k_{osculating\ circle}} = \frac{1}{k_{curve\ (L)}} = \left| \frac{ds}{d\varphi} \right| = \frac{(\dot{x}^2 + \dot{y}^2)^{3/2}}{|\dot{y}\,\ddot{x} - \ddot{y}\,\dot{x}|} = \frac{(1 + y'^2)^{3/2}}{|y''|} \quad (39)$$

A curve (L) is very close to its osculating circle near its point M. On the basis of Taylor's formula (IV.50) we can show that the curve (L) can be regarded as coinciding with its osculating circle
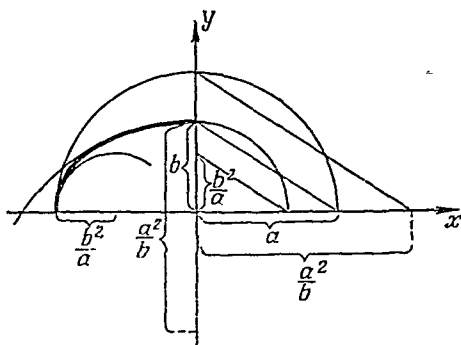
Fig. 179

in the vicinity of the point M with an infinitesimal error of the third order relative to Δt. This assertion, in its turn, implies that the osculating circle may be regarded as a circle passing through three points of the curve (L) lying infinitely close to each other. The last property also applies to curves in space. In particular, it straightway implies that the osculating circle lies in the osculating plane.

The points of a curve at which the curvature assumes its extremal values (but not the points of inflection) are called the *vertices* of the curve. For instance, take the parametric equations $x = a \cos t$ and $y = b \sin t$ of an ellipse. Then

$$k = \frac{|(-b \sin t)(-a \sin t) - (b \cos t)(-a \cos t)|}{[(-a \sin t)^2 + (b \cos t)^2]^{3/2}} = \frac{ab}{(a^2 \sin^2 t + b^2 \cos^2 t)^{3/2}}$$

Here the denominator has extremal values at $t = 0$, $t = \dfrac{\pi}{2}$, $t = \pi$, $t = \dfrac{3}{2}\pi$ etc. (check it up!). Therefore the vertices of an ellipse

understood in the new sense coincide with the vertices defined in Sec. II.10. The radius of curvature is equal to $\frac{b^2}{a}$ at the points of intersection of the ellipse with the x-axis and is equal to $\frac{a^2}{b}$ at the points of intersection with the y-axis. This result is applied to constructing an approximate form of an ellipse: we draw the circles of curvature at the vertices by means of a compass (see Fig. 179) and then connect the circles by the lines drawn with the help of a French curve. For the sake of simplicity only a quarter of the ellipse is shown. The construction of the radii of curvature at the vertices is also demonstrated.

To determine the coordinates ξ and η of the centre of curvature of a curve (L) at a point M let us suppose that the curve is convex downwards at M, that is $y'' > 0$ (see Sec. IV.20), as it is shown in Fig. 177. Then

$$
\left.
\begin{aligned}
\xi &= x - R \sin \varphi = x - R \frac{\tan \varphi}{\sqrt{1+\tan^2 \varphi}} = \\
&= x - \frac{(1+y'^2)^{3/2}}{y''} \frac{y'}{\sqrt{1+y'^2}} = x - \frac{y'(1+y'^2)}{y''} \\
\eta &= y + R \cos \varphi = y + R \frac{1}{\sqrt{1+\tan^2 \varphi}} = \\
&= y + \frac{(1+y'^2)^{3/2}}{y''} \frac{1}{\sqrt{1+y'^2}} = y + \frac{1+y'^2}{y''}
\end{aligned}
\right\} \quad (40)
$$

In case $y'' < 0$ we obtain the same formulas. If we pass to the derivatives with respect to an arbitrary parameter we obtain (check it!)

$$
\xi = x - \frac{\dot{y}(\dot{x}^2+\dot{y}^2)}{\dot{y}\,\ddot{x}-\ddot{y}\,\dot{x}}, \qquad \eta = y + \frac{\dot{x}(\dot{x}^2+\dot{y}^2)}{\dot{y}\,\ddot{x}-\ddot{y}\,\dot{x}} \qquad (41)
$$

26. Evolute and Evolvent.

A curve (L) given, consider the locus of centres of curvature of the curve (L) and denote it by $(\overline{L})$. This locus is called the **evolute of the curve**. In its turn, the curve (L) itself is called the **evolvent (or involute)** with respect to its evolute $(\overline{L})$.

Thus, if $(\overline{L})$ is the evolute of (L) then (L) is the evolvent of $(\overline{L})$ and vice versa. It can be shown that we have the following relationships between an evolute and its evolvent.

(1) Let M be an arbitrary point of the evolvent and $\overline{M}$ the corresponding point of the evolute, i.e. $\overline{M}$ is the centre of curvature of the curve (L) at the point M. Then the straight line $M\overline{M}$ is not

only the normal to the evolvent but also the tangent to the evolute. This relationship is illustrated in Fig. 180.

(2) Let a point M move along the evolvent. Then the increment of the radius of curvature is equal to the length of the corresponding arc of the evolute lying between the centres of curvature (see Fig. 181).

These properties imply the following characteristic feature of an evolvent: if an unstretchable taut thread is wound off the contour having the form of the evolute the end of the thread describes the
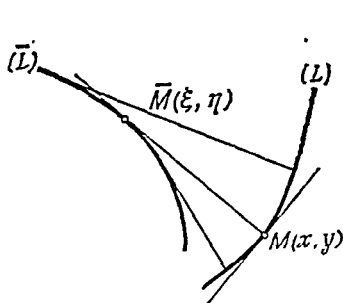
Fig. 180   Fig. 181

$(\bar{L})$ is the evolute, (L) is the evolvent
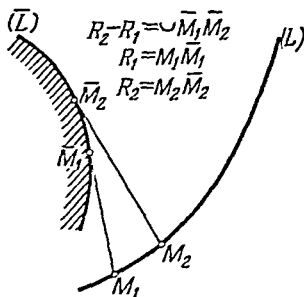
evolvent. The involute (evolvent) of a circle is of practical importance because the profiles of the lateral surfaces of teeth of most gear wheels are shaped in the form of the evolvent of a circle.

If a curve (L) is represented in the parametrical form $x = x(t)$ and $y = y(t)$ then the parametrical equations of its evolute $\xi = \xi(t)$ and $\eta = \eta(t)$ (where ξ and η are the coordinates on the same x- and y-axis) are obtained from formulas (41) by substituting the expressions of x and y in terms of t into the formulas.

To prove the two properties let us first differentiate equalities (40) (which hold if $\frac{d\varphi}{ds} > 0$):

$$d\xi = dx - dR \sin \varphi - R \cos \varphi \, d\varphi,$$
$$d\eta = dy + dR \cos \varphi - R \sin \varphi \, d\varphi \qquad (42)$$

But formula (39) implies

$$dx = ds \cos \varphi = \frac{ds}{d\varphi} \cos \varphi \, d\varphi = R \cos \varphi \, d\varphi$$
$$\text{and, similarly,} \quad dy = R \sin \varphi \, d\varphi \qquad (43)$$

Substituting (43) into (42) we receive

$$d\xi = -dR \sin \varphi, \quad d\eta = dR \cos \varphi \qquad (44)$$

The same final result (44) would be obtained if we considered the case $\frac{d\varphi}{ds} < 0$.

The immediate consequence of formulas (44) is that $\frac{d\eta}{d\xi} =$ $= -\cot \varphi = -\dfrac{1}{\dfrac{dy}{dx}}$. This means (see problem 5 in Sec. II.9) that the tangent to the evolvent at its arbitrary point M and the tangent to the evolute at the corresponding point $\overline{M}$ are mutually perpendicular (see Fig. 180) which is just the first property.

From formulas (44) we also deduce $d\xi^2 + d\eta^2 = dR^2$. This implies (see the end of Sec. 23) $dR = \pm ds$ where s is the arc length reckoned along the evolute. Consequently, $d(R \mp s) = 0$, $R \mp s = \text{const}, R = C \pm s, \Delta R = \pm \Delta s$ and $|\Delta R| = |\Delta s|$. This furnishes the proof of the second property of the evolute and the evolvent.
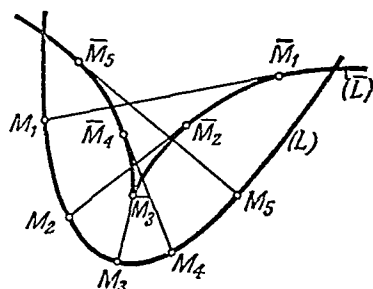
It may be shown that to the vertices of a curve (see the end of Sec. 25) there correspond cusps of its evolute (see Fig. 182). For example, the evolute of an ellipse has four cusps. If a curve has a zero curvature at a point (in particular, such a situation usually occurs at points of inflection) the corresponding point of its evolute travels into infinity.

Fig. 182

As an example, let us determine the evolute of a cycloid [see formulas (II.12) in which we substitute t for ψ]. Here we have

$$x = R(t - \sin t). \quad \dot{x} = R(1 - \cos t), \quad \ddot{x} = R \sin t;$$
$$y = R(1 - \cos t), \quad \dot{y} = R \sin t, \quad \ddot{y} = R \cos t;$$
$$\dot{x}^2 + \dot{y}^2 = R^2(1 - \cos t)^2 + R^2 \sin^2 t = 2R^2(1 - \cos t);$$
$$\dot{y}\ddot{x} - \dot{y}\ddot{x} = R \cos t \cdot R(1 - \cos t) - R \sin t \cdot R \sin t =$$
$$= -R^2(1 - \cos t)$$

By formulas (41) we obtain

$$\xi = R(t - \sin t) - \frac{R \sin t \cdot 2R^2(1 - \cos t)}{-R^2(1 - \cos t)} = R(t + \sin t)$$

and, similarly, $\eta = -R(1 - \cos t)$. Now denoting $t = \pi + \tau$ we receive

$$\xi = R[\pi + \tau + \sin(\pi + \tau)] = R(\tau - \sin \tau) + \pi R,$$
$$\eta = -R[1 - \cos(\pi + \tau)] = R(1 - \cos \tau) - 2R$$

Since πR and $2R$ are constants we see [comparing with formulas (II.12)] that the evolute of a cycloid is the cycloid of the same sizes but translated πR units of length in the positive direction of the x-axis and $2R$ units of length in the negative direction of the y-axis with respect to the original cycloid. This fact is sometimes utilized in engineering.

It also follows from Fig. 181 that an evolvent is a special case of a roulette (see Sec. II.6) described by a point of a straight line when the line rolls upon the evolute. Incidentally, this result shows that the arc of any curve can serve as a roulette, that is roulettes do not form a special class of curves.

Complex Numbers
and Functions

Complex numbers are widely used in modern mathematics and its applications. It turns out that it is convenient to obtain many relationships between real quantities by using complex numbers and functions in intermediate calculations.

§ 1. Complex Numbers

1. Complex Plane. The definition of a complex number is well known from elementary mathematical courses. A complex number is an expression of the form

$$z = x + iy \tag{1}$$

where x and y are real numbers and i is the **imaginary** unit satisfying the equality $i^2 = -1$; x is called the **real part** and y the **imaginary part** of the complex number z. For the real and imaginary parts we shall use the notation $x = \operatorname{Re} z$ and $y = \operatorname{Im} z$ (the term "imaginary part" is sometimes applied to the whole product iy which is more natural but less convenient). *Two complex numbers are equal if and only if their real parts are equal and their imaginary parts are equal*: if $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ then the equalities

$$z_1 = z_2 \quad \text{and} \quad \left. \begin{array}{l} x_1 = x_2 \\ y_1 = y_2 \end{array} \right\} \tag{2}$$

are equivalent. Hence, *one "complex equality" is equivalent to two real equalities. Signs of inequality cannot be applied to complex numbers.* that is inequalities of the form $z_1 > z_2$ do not exist.

Complex numbers may be represented in a plane. To do this we take a Cartesian coordinate system x. y which enables us to represent any number of the form (1) as a point $M\,(x;\,y)$. Such a plane is conventionally called a **complex plane** but, of course, all the points of the plane have real coordinates. For brevity we often say

he point $x + iy$" instead of "the point corresponding to the number
$+ iy$" (see Fig. 183).

Real numbers are a special case of complex numbers: if we put
$= 0$ in formula (1) we may regard a complex number $x + i\cdot0$
;representing the real number x; real numbers are represented as
oints lying on the **real axis (axis of reals)**, i.e. on the x-axis. Com-
lex numbers that are not real are called **imaginary**. Thus, every
omplex number is either real or imaginary. A complex number
ithout a real part (i.e. with the real part equal to zero) is called
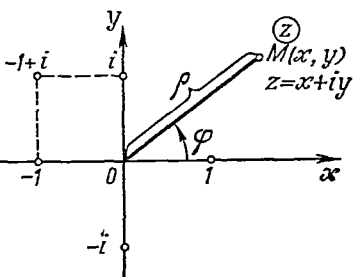
Fig. 183
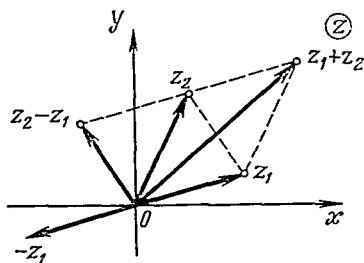
Fig. 184

ure **imaginary**; such numbers are represented by points of the
axis which is called the **imaginary axis** or the **axis of imaginaries**.
queer thing in the terminology is that the number $z = 0$ is a pure
naginary number but not an imaginary number!

It is often convenient to introduce polar coordinates in a complex
ane (see Sec. II.3 and Fig. 183). The polar coordinates ρ and φ
the point $M(x; y)$ which represents the complex number $z =$
$= x + iy$ are denoted as $\rho = |z|$ and $\varphi = \arg z$. ρ is called the
odulus or the absolute value of z and φ is called the argument, or
mplitude, or phase, of z.

As is known, $\rho = \sqrt{x^2 + y^2}$, $x = \rho \cos \varphi$ and $y = \rho \sin \varphi$. This
nplies, by (1), that

$$z = \rho (\cos \varphi + i \sin \varphi) \tag{3}$$

ence, each complex number can be written in the so-called *tri-*
onometric form (3). The modulus of a complex number is a certain
niquely defined non-negative real number whereas the argument
defined within an integral multiple of 2π. For instance, $|i| = 1$
d $\text{Arg } i = \frac{\pi}{2} + 2k\pi$ $(k = 0, \pm1, \pm2, \ldots)$. The sign $\text{Arg } z$
enotes the totality of all the possible values of the argument of
complex number z. Thus, $\text{Arg } z$ has infinitely many different
alues. There is, however, one and only one value of $\text{Arg } z$, denoted
$\arg z$, which satisfies the inequality $-180° < \arg z \leqslant 180°$;

arg z is called the **principal value of the argument**. Besides, any arbitrary value can be taken as the argument of the number $z = 0$.

2. Algebraic Operations on Complex Numbers. To carry out the addition of complex numbers we should sum their real parts and their imaginary parts separately. Comparing this rule with the rule of addition of vectors [see formula (VII.10)] we see that the addition and subtraction of complex numbers are performed in the same way as the operations on vectors (see Fig. 184). In particular, it follows that

$$| z_1 + z_2 | \leqslant | z_1 | + | z_2 |$$

Similarly, it can be verified that complex numbers are multiplied by real numbers in the same way as vectors. These properties make it possible to interpret complex numbers as vectors in the complex plane. The connection between the representation of complex numbers as points of the complex plane and their vector representation is obvious: if the vector is drawn from the origin of the coordinate system its terminus is at the corresponding point representing the complex number.

The rule of multiplication of complex numbers is quite different from that of vectors. If we use the trigonometric form

$$z_1 = \rho_1 (\cos \varphi_1 + i \sin \varphi_1), \quad z_2 = \rho_2 (\cos \varphi_2 + i \sin \varphi_2)$$

then the product $z = z_1 \cdot z_2$ can be written as

$$z = \rho (\cos \varphi + i \sin \varphi) = \rho_1 (\cos \varphi_1 + i \sin \varphi_1) \rho_2 (\cos \varphi_2 + $$
$$+ i \sin \varphi_2) = \rho_1 \rho_2 (\cos \varphi_1 \cos \varphi_2 + i \cos \varphi_1 \sin \varphi_2 + $$
$$+ i \sin \varphi_1 \cos \varphi_2 - \sin \varphi_1 \sin \varphi_2) = $$
$$= \rho_1 \rho_2 [\cos (\varphi_1 + \varphi_2) + i \sin (\varphi_1 + \varphi_2)]$$

Therefore,

$$\rho = \rho_1 \rho_2, \quad \varphi = \varphi_1 + \varphi_2$$

i.e.

$$| z_1 z_2 | = | z_1 | \cdot | z_2 |, \quad \text{Arg} (z_1 \cdot z_2) = \text{Arg} \, z_1 + \text{Arg} \, z_2$$

Hence, *when complex numbers are multiplied their moduli are mul tiplied and their arguments are added.* This implies that for the inverse operation, the division, we have

$$\left| \frac{z_1}{z_2} \right| = \frac{z_1 |}{| z_2|}, \quad \text{Arg} \frac{z_1}{z_2} = \text{Arg} \, z_1 - \text{Arg} \, z_2$$

The multiplication of a complex number by i is of special interest: $| iz | = | z |$ and $\text{Arg} \, iz = \text{Arg} \, z + \frac{\pi}{2}$ since $| i | = 1$ and $\arg i = \frac{\pi}{2}$; thus, the vector iz is obtained by turning the vector z in the positive direction through a right angle.

The rule of multiplication of complex numbers extends automatically to an arbitrary number of factors. In particular, if we take equal factors then we have

$$[\rho \, (\cos \varphi + i \sin \varphi)]^n = \rho^n \, (\cos n\varphi + i \sin n\varphi)$$
$$(n = 2, \, 3, \, \ldots)$$

In case $\rho = 1$ we obtain the formula

$$(\cos \varphi + i \sin \varphi)^n = \cos n\varphi + i \sin n\varphi$$

which was named **De Moivre's formula** after the English mathematician A. Moivre (1667-1754) who discovered the formula in 1707.

De Moivre's formula can be applied to express the values of trigonometric functions of multiple arcs. For example, taking $n = 3$ we get

$$(\cos \varphi + i \sin \varphi)^3 = \cos^3 \varphi + i \, 3 \cos^2 \varphi \sin \varphi - $$
$$- 3 \cos \varphi \sin^2 \varphi - i \sin^3 \varphi = \cos 3\varphi + i \sin 3\varphi$$

which implies, by formula (2), that

$$\cos 3\varphi = \cos^3 \varphi - 3 \cos \varphi \sin^2 \varphi, \quad \sin 3\varphi =$$
$$= 3 \cos^2 \varphi \sin \varphi - \sin^3 \varphi$$

Here, of course, we should take into account the table of powers of the number i: $i^1 = i$, $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, $i^5 = i$, $i^6 = -1$ etc. By the way, note that $\frac{1}{i} = -i$.

Now we turn to extracting roots of complex numbers. If $z = \rho \, (\cos \varphi + i \sin \varphi)$ is given and $\sqrt[n]{z} = w = r \, (\cos \psi + i \sin \psi)$ is to be found. then, by the definition of a root, $z = w^n = r^n \, (\cos n\psi + i \sin n\psi)$. Comparing this with the original expression of z we conclude (see the end of Sec. 1) that

$$r^n = \rho, \quad n\psi = \varphi + 2k\pi \quad (k \text{ is an arbitrary integer})$$

Since r and ρ are non-negative numbers, we have $r = (\sqrt[n]{\rho})_{ord}$ and $\psi = \frac{\varphi + 2k\pi}{n}$ where $(\sqrt[n]{\rho})_{ord}$ denotes the "ordinary" (i.e. the arithmetic. real positive) root of a non-negative number. Thus,

$$w = (\sqrt[n]{\rho})_{ord} \left(\cos \frac{\varphi + 2k\pi}{n} + i \sin \frac{\varphi + 2k\pi}{n} \right)$$

Making k assume the values 0, 1, 2, ... we shall get all the values w_1, w_2, w_3, ... of the root. But for $k = n$ we obtain

$$w_{n+1} = \sqrt[n]{\rho})_{ord} \left(\cos \frac{\varphi + 2n\pi}{n} + i \sin \frac{\varphi + 2n\pi}{n} \right) =$$
$$= (\sqrt[n]{\rho})_{ord} \left[\cos \left(\frac{\varphi}{n} + 2\pi \right) + i \sin \left(\frac{\varphi}{n} + 2\pi \right) \right] =$$
$$= (\sqrt[n]{\rho})_{ord} \left(\cos \frac{\varphi}{n} + i \sin \frac{\varphi}{n} \right) = w_1$$

Similarly, $w_{n+2} = w_2$ etc. For negative k we do not obtain any new values either: $k = -1$ yields the same result as $k = n - 1$ etc. Finally,

$$[\sqrt[n]{\rho\,(\cos\varphi + i\sin\varphi)}]_{k+1} =$$
$$= (\sqrt[n]{\bar\rho})_{ord}\left(\cos\frac{\varphi + 2k\pi}{n} + i\sin\frac{\varphi + 2k\pi}{n}\right) \quad (k = 0,\,1,\,\ldots,\,n-1) \qquad (4)$$

We see now that the nth root of a complex number has n different values; the number $z = 0$ makes the only exception to this rule since all the roots of 0 are equal to zero.

For example, since $2i = 2\left(\cos\dfrac{\pi}{2} + i\sin\dfrac{\pi}{2}\right)$, we have

$$(\sqrt{2i})_{1,2} = (\sqrt{2})_{ord}\left(\cos\frac{\dfrac{\pi}{2} + 2k\pi}{2} + i\sin\frac{\dfrac{\pi}{2} + 2k\pi}{2}\right) \quad (k = 0,\,1)$$

which enables us to calculate $(\sqrt{2i})_1 = 1 + i$ and $(\sqrt{2i})_2 = -1 - i$ easily.

Consider another example:

$$(\sqrt[3]{1})_{1,2,3} = (\sqrt[3]{1})_{ord}\left(\cos\frac{2k\pi}{3} + i\sin\frac{2k\pi}{3}\right) \quad (k = 0,\,1,\,2)$$

since $1 = 1(\cos 0 + i\sin 0)$. This implies $(\sqrt[3]{1})_1 = 1$, $(\sqrt[3]{1})_2 = -\dfrac{1}{2} + i\dfrac{\sqrt{3}}{2}$ and $(\sqrt[3]{1})_3 = -\dfrac{1}{2} - i\dfrac{\sqrt{3}}{2}$. In this case one of the roots has turned out to be ordinary, real whereas the other two roots are imaginary.

The geometrical meaning of formula (4) is illustrated in Fig. 185 where we have taken $n = 5$.

3. **Conjugate Complex Numbers.** The number $z^* = x - iy$ is called conjugate to the number $z = x + iy$; we often write $\bar z$ instead of z^*. The simple properties of conjugate numbers are the following:

(1) $(z^*)^* = (x - iy)^* = [x + i\,(-y)]^* = x - i\,(-y) = x + iy = z$, i.e. the numbers z and z^* are mutually conjugate;

Fig. 185

(2) $z + z^* = 2\mathrm{Re}\,z$, $z - z^* = 2i\,\mathrm{Im}\,z$;

(3) $z^* = z$ if and only if z is real;

(4) $zz^* = (x - iy)(x + iy) = x^2 + y^2 = |z|^2$;

(5) $|z^*| = |z|$, $\mathrm{Arg}\,z^* = -\mathrm{Arg}\,z$, i.e. the points z and z^* are symmetric with respect to the real axis;
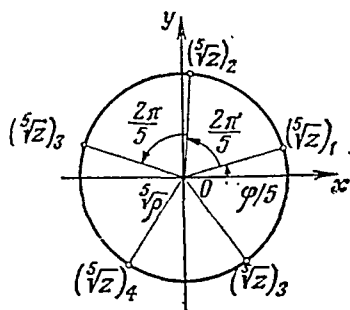
(6) $(z_1 + z_2)^* = z_1^* + z_2^*$ since

$$(z_1 + z_2)^* = (x_1 + iy_1 + x_2 + iy_2)^* =$$
$$= [x_1 + x_2 + i\,(y_1 + y_2)]^* = x_1 + x_2 - i\,(y_1 + y_2) =$$
$$= (x_1 - iy_1) + (x_2 - iy_2) = z_1^* + z_2^*$$

(7) $(z_1 z_2)^* = z_1^* z_2^*$ which can be verified in the same way as property (6).

If we substitute $\dfrac{z_1}{z_2}$ for z_1 into the formula expressing property (7) we get $z_1^* = \left(\dfrac{z_1}{z_2}\right)^* z_2^*$ which yields

(8) $\left(\dfrac{z_1}{z_2}\right)^* = \dfrac{z_1^*}{z_2^*}$.

Properties (6) and (7) extend automatically to any arbitrary number of summands or factors. For instance,

$$(z^n)^* = (z^*)^n,$$
$$(2z^n + iz^m)^* = (2z^n)^* + (iz^m)^* = 2\,(z^*)^n - i\,(z^*)^m \quad \text{etc.}$$

Generally, to pass from any rational expression containing an arbitrary number of variables and coefficients to the conjugate expression it is necessary to replace each variable and each coefficient by its conjugate value. It can be shown that this rule holds not only for rational expressions but also for irrational expressions, for sums of power series and so on. It follows that each equality involving complex expressions of the above type remains true when $-i$ is substituted for i everywhere in the equality because the substitution transforms the original equality of complex numbers into the equality of their conjugate numbers. The numbers i and $-i$ are therefore indistinguishable in the algebraic sense. It would be incorrect to say that $i = \sqrt{-1}$ and $-i = -\sqrt{-1}$. In fact the root $\sqrt{-1}$ simply has two different values designated as $\pm i$.

Conjugate numbers can be used, in particular, to separate the real and the imaginary parts of a fraction of the form $\dfrac{z_1}{z_2} = \dfrac{x_1 + iy_1}{x_2 + iy_2}$. To do this we multiply the numerator and the denominator by z_2^* and obtain the fraction with the real denominator which enables us to perform the separation easily.

For example,

$$\operatorname{Re}\frac{2 + i5}{3 - i2} = \operatorname{Re}\frac{(2 + i5)\,(3 + i2)}{(3 - i2)\,(3 + i2)} = \operatorname{Re}\frac{6 + i4 + i15 - 10}{13} =$$
$$= \operatorname{Re}\left(-\frac{4}{13} + i\,\frac{19}{13}\right) = -\frac{4}{13}$$

4. Euler's Formula. Now we turn to the transcendental operations on complex numbers. In Sec. IV.16 we showed that for real x,

$$e^x = 1 + \frac{1}{1!}\,x + \frac{1}{2!}\,x^2 + \frac{1}{3!}\,x^3 + \ldots \tag{5}$$

If we substitute z for x we shall come to *the definition of the exponential function with a complex exponent*: it is defined as

$$e^z = 1 + \frac{1}{1!}z + \frac{1}{2!}z^2 + \frac{1}{3!}z^3 + \cdots \qquad (6)$$

We shall show in Sec. XVII.14 that this definition can be justified for all z and besides the basic property of the exponential function remains true:

$$e^{z_1}e^{z_2} = e^{z_1+z_2} \qquad (7)$$

Formulas (6) and (5) show that in the special case when z is real the new definition of e^z coincides with the old one; in general, each new definition must not contradict the facts already established. At the same time formula (7) confirms the expedience of this definition of e^z.

In the same standard way we can define, for the complex values of the argument, functions $f(x)$ which were originally defined for real values of the argument. For this purpose we should expand a given function $f(x)$ into Taylor's series in powers of x or in powers of $x - a$ where a is a real number and then replace x by z and denote the sum of the series by $f(z)$. Thus, by analogy with (6), we write, using formulas (IV.56) and (IV.57), for complex z,

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots \qquad (8)$$

$$\cos z = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \cdots \qquad (9)$$

etc. In Sec. XVII.14 we shall show that all the basic formulas which have the form of identical equalities and hold for real values of the argument (for instance, such as $\sin(-x) \equiv -\sin x$ or $\sin^2 x + \cos^2 x \equiv 1$ etc.) remain true for the complex values of the argument.

The above formulas reveal the essential relationship between the exponential function and trigonometric functions. Namely, substituting iz for z into (6) we deduce

$$e^{iz} = \left(1 + \frac{1}{1!}iz - \frac{1}{2!}z^2 - \frac{1}{3!}iz^3 + \frac{1}{4!}z^4 + \frac{1}{5!}iz^5 - \frac{1}{6!}z^6 - \right.$$
$$\left. - \frac{1}{7!}iz^7 + \cdots\right) = \left(1 - \frac{1}{2!}z^2 + \frac{1}{4!}z^4 - \frac{1}{6!}z^6 + \cdots\right) +$$
$$+ i\left(\frac{1}{1!}z - \frac{1}{3!}z^3 + \frac{1}{5!}z^5 - \frac{1}{7!}z^7 + \cdots\right)$$

This, together with (8) and (9), implies **Euler's formula**

$$e^{iz} = \cos z + i \sin z \qquad (10)$$

which is very important. The formula

$$e^{-iz} = e^{i(-z)} = \cos(-z) + i \sin(-z) = \cos z - i \sin z$$

and the formulas

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} \quad \text{and} \quad \sin z = \frac{e^{iz} - e^{-iz}}{2i} \qquad (11)$$

which are implied by the above formula and (10) are also often used. All the formulas were discovered by Euler in 1743.

From Euler's formula (10), using property (7), we obtain the following expression for the exponential function with an arbitrary complex exponent:

$$e^z = e^{x+iy} = e^x e^{iy} = e^x (\cos y + i \sin y) \qquad (12)$$

The comparison with trigonometric form (3) shows that

$$|e^z| = e^x, \quad \operatorname{Arg} e^z = y + 2k\pi \qquad (13)$$

In particular, it is seen that we always have $|e^z| > 0$, i.e. $e^z \neq 0$. If we write z instead of e^z in formula (12) then, by (13), we obtain

$$z = |z| (\cos \arg z + i \sin \arg z) = |z| e^{i \arg z} = \rho e^{i\varphi}$$

This "exponential form" of complex numbers is convenient for performing algebraic operations on them.

Formulas (11) imply the following relationships between trigonometric and hyperbolic functions (see Sec. I.28): $\cos z = \cosh iz$ and $\sin z = \frac{\sinh iz}{i}$, that is $\sinh iz = i \sin z$. From this, substituting iz for z, we also deduce $\cos iz = \cosh z$ and $\sin iz = i \sinh z$.

It is these formulas that reveal the essential relationship between the functions (the relationship was mentioned in Sec. I.28) which enables us to transform relations between trigonometric functions into the corresponding relations between hyperbolic functions and vice versa. (Let the reader deduce the basic relation between $\cosh z$ and $\sinh z$ by substituting iz for z into the formula $\cos^2 z + \sin^2 z = 1$.)

Using formulas (11) we can also obtain the expression of powers of sine and of cosine in terms of trigonometric functions of multiple arguments. For instance,

$$\cos^3 x = \left(\frac{e^{ix} + e^{-ix}}{2} \right)^3 = \frac{e^{i \cdot 3x} + 3e^{ix} + 3e^{-ix} + e^{-i \cdot 3x}}{8} =$$
$$= \frac{e^{i \cdot 3x} + e^{-i \cdot 3x}}{8} + \frac{3}{8} (e^{ix} + e^{-ix}) = \frac{\cos 3x}{4} + \frac{3 \cos x}{4} \qquad (14)$$

etc. Transformations of this kind are used for integrating trigonometric functions.

5. Logarithms of Complex Numbers. The definition of "complex logarithms" is essentially the same as that for logarithms of real

numbers: the logarithm of a complex number z is the number w for which $z = e^w$. To find the value of the logarithm let us denote

$$z = \rho (\cos \varphi + i \sin \varphi), \quad w = u + iv$$

Then we deduce from formula (12):

$$\rho (\cos \varphi + i \sin \varphi) = z = e^w = e^u (\cos v + i \sin v)$$

Since u and v are real this implies

$$e^u = \rho, \quad \text{i.e.} \quad u = \ln \rho; \quad v = \varphi + 2k\pi \quad (k \text{ is an integer})$$

where $\ln \rho$ is understood as an "ordinary", real logarithm of a positive number. Thus,

$$\text{Ln } z = w = u + iv = \ln \rho + i\varphi + i2k\pi = \ln |z| + i \text{ Arg } z$$

where Ln z denotes the totality of all the values of the logarithm.

Hence, a logarithm of a complex number has an infinite set of different values. The number "zero" is the only exception to the rule because it has no logarithm. We can write conditionally (formally): $\text{Ln } 0 = -\infty + iv$ where v is arbitrary.

Real positive numbers being a special case of complex numbers, their logarithms are also infinite-valued. One of the values is "ordinary", real, whereas all the others are imaginary. For example,

$$\text{Ln } 1 = \ln 1 + i0 + i2k\pi = i2k\pi \quad (k = 0, \pm 1, \pm 2, \ldots)$$

For $k = 0$ we get the original value $\ln 1 = 0$ but at the same time we can take for the logarithm of 1 the values $i2\pi$, $-i2\pi$, $i4\pi$ etc. Let us check it up once again:

$$e^{i2k\pi} = \cos 2k\pi + i \sin 2k\pi = 1 + i0 = 1 \qquad (15)$$
$$(k = 0, \pm 1, \pm 2, \ldots)$$

Negative real numbers also have logarithms but all their values are imaginary. For example, $\text{Ln } (-1) = i\pi (2k + 1)$ (check it up!).

By means of logarithms we raise a complex number to an arbitrary complex power: the involution is defined as

$$z_1^{z_2} = (e^{\text{Ln } z_1})^{z_2} = e^{z_2 \text{Ln } z_1}$$

and the right-hand side is calculated by formula (12). Since logarithms are infinite-valued the whole power is also infinite-valued in the general case.

§ 2. Complex Functions of a Real Argument

6. Definition and Properties. It is sometimes necessary to deal with functions which assume complex values although their independent variable, the argument, is real. As examples of such functions we can take

(1) $z = (t + i)^3$; (2) $z = Me^{pt}$ $(p = a + i\omega)$ etc.

Here the independent variable is denoted by t and the functions are denoted by z. If we decompose the value of such a function into its real and imaginary parts, i.e. $z = x + iy$, then each of the parts will be a function of t; thus, in the above examples we obtain

(1) $x = t^3 - 3t, \quad y = 3t^2 - 1;$ (2) $x = Me^{at} \cos \omega t,$
$$y = Me^{at} \sin \omega t$$

(verify these relations taking into account that M is real!).

In the general case, if

$$z = f(t) = \varphi(t) + i\psi(t) \qquad (16)$$

we obtain

$$x = \varphi(t), \qquad y = \psi(t) \qquad (17)$$

Conversely, (17) yields (16). The indication of a complex function of a real argument is therefore equivalent to the indication of two ordinary, real, functions of the argument. This situation is quite analogous to the case when we have a vector function of a scalar argument (see Sec. VII.23). The analogy becomes still more complete if we interpret complex numbers as vectors (see Sec. 2).

It follows that the theory of complex functions of a real argument does not involve any essentially new features in comparison with the theory of real functions. In particular, the definitions of the continuity, of the derivative etc. are transferred without any changes. From the considerations in Sec. XVII.14 it follows that all the differentiation formulas remain true. For instance,

$$[(t + i)^3]' = 3(t + i)^2, \quad (Me^{pt})' = Mpe^{pt} \quad \text{and so forth}$$

A function of form (16) is represented by a curve in a complex plane which has parametric equations (17).

When using functions of form (16) the following obvious properties should be taken into account:

if complex functions are added up their real parts and their imaginary parts are added separately;

if a complex function is multiplied by a real constant or by a real function the real and the imaginary parts are also multiplied by the same factor;

if a complex function is differentiated the same operation is performed on its real and imaginary parts.

The properties are expressed by the formulas

$$\text{Re } [f_1(t) + f_2(t)] = \text{Re } f_1(t) + \text{Re } f_2(t) \quad \text{etc.}$$

(Deduce the formulas!)

These properties make it possible to perform the above operations on the whole complex function and then to take the real or

the imaginary part of the resulting expression instead of performing the operations on the real or on the imaginary part. It is remarkable that such a transition to complex quantities together with the reverse transition to the real quantities which are sought for sometimes turns out to be more obvious and simpler than the corresponding direct operations on the real quantities.

Here we also mention a function of the form

$$\mathrm{Ln}\,[t - (a + ib)] = \ln|t - a - ib| + i\,\mathrm{Arg}\,[t - a - ib] =$$
$$= \frac{1}{2}\ln[(t-a)^2 + b^2] + i\left[\text{arc}\,\tan\frac{t-a}{b} - \frac{\pi}{2} + k\pi\right]$$

The function is sometimes of use for certain applications when the integer k is chosen in an appropriate way.

7. Applications to Describing Oscillations. Let us take the function

$$U(t) = Me^{i\,(\omega t + \alpha)} = M\cos(\omega t + \alpha) + iM\sin(\omega t + \alpha)$$
$$(M > 0,\ \omega > 0) \tag{18}$$

This function is convenient to apply for investigating harmonic oscillations (compare with Sec. I.29). For this purpose it should be noted that expression (18) has the modulus M and the argument $\omega t + \alpha$, i.e. it is represented by a vector of constant length which rotates uniformly with the angular speed ω.

For example, let us consider the superposition of oscillations with equal frequencies. Let it be necessary to sum up the two quantities $u_1(t) = M_1\sin(\omega t + \alpha_1)$ and $u_2(t) = M_2\sin(\omega t + \alpha_2)$. To do this we introduce the corresponding complex quantities $U_1(t) = M_1 e^{i\,(\omega t + \alpha_1)}$ and $U_2(t) = M_2 e^{i\,(\omega t + \alpha_2)}$, u_1 and u_2 being, respectively, their imaginary parts. The vectors $U_1(t)$ and $U_2(t)$ rotate uniformly with the angular speed ω and therefore the vector $U_1(t) + U_2(t)$ rotates uniformly with the same speed. Hence this vector can be represented in form (18). To find M and α it is sufficient to consider the situation at the moment $t = 0$ (see Fig. 186). The figure shows that projecting on the coordinate axes we obtain

$$\left.\begin{array}{l} M\cos\alpha = M_1\cos\alpha_1 + M_2\cos\alpha_2 \\ M\sin\alpha = M_1\sin\alpha_1 + M_2\sin\alpha_2 \end{array}\right\} \tag{19}$$

Taking the imaginary part of $U(t)$ we finally conclude that $u_1(t) + u_2(t) = M\sin(\omega t + \alpha)$ where M and α should be found from equalities (19) (find them and explain the corresponding formula for M by means of Fig. 186).

A similar result follows when we have the superposition of an arbitrary number of harmonic oscillations with the same frequency. The superposition of oscillations having different frequencies will be discussed in Sec. XVII.23.

The advantage of exponential function (18) over the trigonometric functions is particularly well seen when we differentiate:

$$\frac{dU}{dt} = Me^{i(\omega t+\alpha)}i\omega = i\omega U \tag{20}$$

By Sec. 2, we again obtain a vector which rotates uniformly with the angular speed ω but it leads U by 90° and has a modulus which is ω times that of U. The rotation and the stretching are repeated when the differentiation of U is continued.

Now we are going to demonstrate an application of functions of form (18) to an electric circuit shown in Fig. 187. There is a resistance R and an inductance L in the circuit. If an alternating
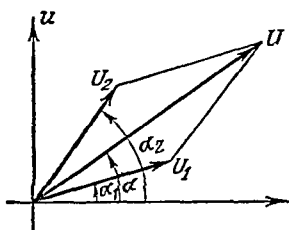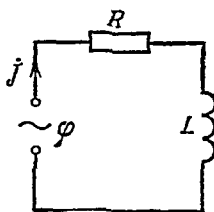
Fig. 186   Fig. 187

potential difference which varies according to the law $\varphi = \varphi_0 \sin(\omega t + \beta)$ is applied to the terminals of the circuit there appears a steady-state alternating electric current flow also varying harmonically: $j = j_0 \sin(\omega t + \alpha)$ but j_0 and α are not known beforehand. Equalling φ to the resulting voltage drop on R and L we deduce, using the well-known physical laws, the basic equation of our problem: $Rj + L\frac{dj}{dt} = \varphi$.

Let us introduce the notion of a *complex voltage* and that of a *complex current* by formulas $\Phi = \varphi_0 e^{i(\omega t+\beta)}$ and $J = j_0 e^{i(\omega t+\alpha)}$. The "real" voltage and current are equal to the imaginary parts of the corresponding expressions. By the properties described in Sec. 6, to find j we must solve the equation $RJ + L\frac{dJ}{dt} = \Phi$ and then take the imaginary part of the resulting expression. According to formula (20) we receive

$$RJ + Li\omega J = \Phi, \quad \text{i.e.} \quad J = \frac{\Phi}{R + i\omega L} \tag{21}$$

We see that the inductance L can be interpreted as a certain resistance equal to $i\omega L$; this quantity is called the *impedance* of the unit L. Writing the expression $(R + i\omega L)^{-1}$ in the exponential form

$(R + i\omega L)^{-1} = re^{-i\theta}$ we deduce from (21) (verify it!):

$$j_0 = r\varphi_0 = \varphi_0 \, (R^2 + \omega^2 L^2)^{-\frac{1}{2}}, \qquad \alpha = \beta - \text{arc tan} \, \frac{\omega L}{R} \qquad (22)$$

(Let the reader try to derive the expression for the vector $j_0 e^{i\alpha} = J \,|_{t=0}$ by using the first equality (21) and the geometric method, and then deduce formula (22) from the expression.)

§ 3. The Concept of a Function of a Complex Variable

The theory and the applications of complex functions of a complex variable contain many new ideas and facts in comparison with the theory of functions of a real argument. The theory is dealt with in many books. To a beginner we should recommend [15], [40], [44. Vol. 3, Part 1]. Here we shall give only some simple facts which are directly related to the subject matter of our course.

8. Factorization of a Polynomial. Let us take an entire rational function of the argument $z = x + iy$. that is a polynomial

$$P(z) = a_0 z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n \qquad (a_0 \neq 0) \qquad (23)$$

of the nth degree with certain coefficients $a_0, \ldots, a_n$ which can be complex in the general case. In the books mentioned above the reader can find the proof of a remarkable theorem, namely the "fundamental theorem of algebra" which asserts that *every polynomial of degree $n \gg 1$ has at least one complex zero*, i.e. there exists a root of the equation

$$P(z) = 0 \qquad (24)$$

which can be either real or imaginary. (The theorem was proved by D'Alembert in the middle of the 18th century. A more rigorous proof was given by Gauss at the end of the 18th century.) If we denote one of the roots by z_1 then, as it is proved in elementary courses on algebra, $P(z)$ is divisible by the binomial $z - z_1$, that is $P(z) = (z - z_1) P_1(z)$ where $P_1(z)$ is a polynomial of the $(n-1)$th degree. If we repeat the argument for $P_1(z)$ we shall get $P_1(z) = (z - z_2) P_2(z)$. i.e. $P(z) = (z - z_1)(z - z_2) P_2(z)$ where $P_2(z)$ is a polynomial of degree $n - 2$. These considerations can be continued up to the "polynomial of degree zero". i.e. a constant, and thus we receive

$$P(z) = a_0 (z - z_1)(z - z_2) \ldots (z - z_n) \qquad (25)$$

This formula shows that all the numbers $z_1, z_2, \ldots, z_n$ are zeros of the polynomial $P(z)$ and that it has no other zeros.

Thus, an algebraic equation of form (24) of the nth degree has exactly n roots.

Some of the roots of equation (24) may coincide, i.e. they may be repeated. Such roots are called **multiple** (double, i.e. of multipli-

·city 2, or triple, i.e. of multiplicity 3 etc.) in contrast to **simple** roots which are not repeated. When figuring the number of roots we should reckon each root according to its degree of multiplicity.

A simple test for distinguishing the multiplicity of a root is implied by Taylor's formula for a polynomial (IV.46) which, as we shall show in Sec. 11, remains true for a polynomial in a complex variable. Suppose $z = a$ is a root of equation (24). Then $P(a) = 0$ and the formula implies that if

$$P(a) = P'(a) = \ldots = P^{(k-1)}(a) = 0, \quad P^{(k)}(a) \neq 0 \qquad (26)$$

then the polynomial $P(z)$ is divisible by $(z - a)^k$ and is indivisible by $z - a$ to any power higher than k. The value $z = a$ is therefore a root of equation (24) of multiplicity k.

If $P(z)$ is an arbitrary function, even a transcendental one, and relations (26) hold for a value $z = a$ then $z = a$ is also called a root (a zero) of equation (24) of multiplicity k. (In particular. if $P(a) = 0$ and $P'(a) \neq 0$ the value $z = a$ is a simple root.) In the case of an arbitrary $P(z)$ we conclude, by Taylor's series (IV.53), that for a k-tuple root $z = a$ the ratio $P(z)/(z - a)^k$ has a finite limit, as $z \to a$, which does not equal zero. Hence, the ratio has a removable discontinuity at $z = a$ (see Sec. III.13). Thus we can write

$$P(z) = (z - a)^k \frac{P(z)}{(z - a)^k} = (z - a)^k Q(z)$$

where the function $Q(z)$ remains continuous for $z = a$ and $Q(a) \neq 0$.

For example, the value $x = 0$ is a triple root of the equation

$$x - \sin x = 0$$

since $(x - \sin x)|_{x=0} = 0$, $(1 - \cos x)|_{x=0} = 0$, $(\sin x)|_{x=0} = 0$, and $(\cos x)|_{x=0} = 1$.

If we combine equal factors in the factorization (25) we obtain

$$P(z) = a_0 (z - z_1)^{\alpha_1} \ldots (z - z_k)^{\alpha_k} \qquad (27)$$

where $z_1, \ldots, z_k$ are all the pairwise different roots of equation (24) and $\alpha_1, \ldots, \alpha_k$ are their multiplicities. Factorizations (25) and (27) hold for polynomials (23) with real and complex coefficients as well.

If polynomial (23) has real coefficients then, together with every complex root, it has the conjugate root of the same multiplicity. Indeed, if $P(z_m) = 0$ then $[P(z_m)]^* = 0^* = 0$. But, by Sec. 3,

$$[P(z_m)]^* = [a_0 z_m^n + \ldots + a_n]^* = a_0^* (z_m^*)^n + \ldots + a_n^* =$$
$$= a_0 (z_m^*)^n + \ldots + a_n = P(z_m^*)$$

from which it follows that $z = z_m^*$ is also a solution of equation (24). In case $z = z_m$ is a double root of equation (24) we have, in addition, $P'(z_m) = 0$ which implies, in a similar way, that $P'(z_m^*) = 0$ and so on.

Combining the factors which correspond to a pair of mutually conjugate roots of the form $\alpha \pm i\beta$ in factorization (27) we obtain

$$[z - (\alpha + i\beta)] [z - (\alpha - i\beta)] = (z - \alpha)^2 + \beta^2 =$$
$$= z^2 + pz + q \quad (p = -2\alpha, \quad q = \alpha^2 + \beta^2) \tag{28}$$

Such combinations are used for factoring polynomials with real coefficients depending on real argument. It is natural to denote the independent variable as x, and thus we obtain from (27) the expression

$$P(x) = a_0 (x - x_1)^{\alpha_1} \ldots (x - x_r)^{\alpha_r} (x^2 + p_1 x + q_1)^{\beta_1} \ldots$$
$$\ldots (x^2 + p_s x + q_s)^{\beta_s} \tag{29}$$

where the first r parentheses correspond to the real roots and the last s parentheses correspond to s pairs of conjugate imaginary roots. Since p and q are real we conclude that every real polynomial (i.e. a polynomial with real coefficients) can be factored into real linear and quadratic factors. If all the roots are real there are only linear factors in factorization (29) and if all the roots are imaginary then there are only quadratic factors in it. The exponents $\alpha_1, \ldots, \alpha_r$, $\beta_1, \ldots, \beta_s$ are equal to the multiplicities of the corresponding roots; in particular, they equal unity for simple roots.

9. Numerical Methods of Solving Algebraic Equations. To realize factorizations (27) and (29) it is necessary to solve equation (24), that is an algebraic equation of the nth degree. The solution for $n = 2$ is well known from elementary mathematical courses. In textbooks on higher algebra there are formulas for the solutions for $n = 3$ and $n = 4$ which were discovered as early as the 16th century. But the formulas are so complicated that they are almost never used for practical purposes, especially for $n = 4$. In case $n > 4$ there are no general formulas which express solutions in terms of the coefficients of the equation by means of algebraic operations on the coefficients. The non-existence of such formulas was proved by Abel and by É. Galois (1811-1832), a French mathematician, who created the fundamentals· of modern algebra.

But algebraic equations can be solved approximately to within any degree of accuracy! In § V.1 we described some methods of calculating real roots of equations of form $f(x) = 0$. To find imaginary roots we can use Newton's method [see formula (V.7) in which we may regard x as an imaginary quantity] or an· iterative scheme (see Sec. V.3). It is also possible to substitute $z = x + iy$ into equation (24) and then separate the real and the imaginary

parts:

$$P\,(z) = P\,(x + iy) = Q\,(x,\ y) + iR\,(x,\ y)$$

This reduces equation (24) to a system of equations of the form

$$\left. \begin{array}{l} Q\,(x,\ y) = 0 \\ R\,(x,\ y) = 0 \end{array} \right\}$$

which can be solved with the help of methods discussed in Sec. XII.12. These methods are also described in [3], [6], [10], [33] and [42].

There exist methods that are only applicable to algebraic equations. We shall give here a method which was introduced by several authors and, among them, by N. I. Lobachevsky in 1834.

Every algebraic equation can be written in the form

$$z^n + a_1 z^{n-1} + a_2 z^{n-2} + \ldots + a_n = 0 \qquad (30)$$

with the coefficient in the highest power equal to unity (why is it so?). Separating the even and the odd powers we get

$$z^n + a_2 z^{n-2} + \ldots = -a_1 z^{n-1} - a_3 z^{n-3} - \ldots$$

If now we square both sides of the last equality we obtain an equation containing only even powers of z. Therefore, denoting $z^2 = p$ we receive an equation of the nth degree for p (why is it so?) whose roots are the squares of the roots of equation (30). If we then transform the equation in like manner and put $p^2 = q$ we shall arrive at an equation of the nth degree with the roots equal to the fourth powers of the roots of equation (30) etc.

After several transformations of this type the roots with the greatest moduli become the most important. For instance, if equation (30) has the roots $z_1 = 2$, $z_2 = -1$ and $z_3 = \frac{1}{2}$ the next equation has the roots $p_1 = 4$, $p_2 = 1$ and $p_3 = \frac{1}{4}$. The equation following the above equations will have the roots $q_1 = 16$, $q_2 = 1$ and $q_3 = \frac{1}{16}$ etc. After m transformations are carried out we arrive at an equation of the form

$$v^n + C_1 v^{n-1} + \ldots + C_n = 0 \qquad (31)$$

and its roots, yet unknown, are equal to $v_1 = z_1^{2^m}, \ldots, v_n = z_n^{2^m}$. Therefore, by (25), equation (31) has the form

$$(v - z_1^{2^m}) \ldots (v - z_n^{2^m}) = 0$$

Now let us substitute $v = 10^l \overline{v}$ into equation (31) and let us choose the integer l in such a way that in the resulting equation

for $\bar{v}$, after the division by the highest coefficient 10^{nl}, the greatest of the moduli of the coefficients in $\bar{v}^{n-1}$, $\bar{v}^{n-2}$, ..., 1 should be of the order of 1, that is the greatest modulus is neither too large nor too small. Then, since the equation thus obtained, after factoring, must have the form

$$(\bar{v} - \bar{v_1}) \ldots (\bar{v} - \bar{v_n}) = 0 \quad (\text{where } \bar{v_1} = 10^{-l}v_1 =$$
$$= 10^{-l}z_1^{2^m} \text{ etc.)} \tag{32}$$

we conclude that the greatest root (or roots) of equation (32) is (are) of the order of 1 whereas all the other roots are negligibly small. Omitting these roots, i.e. equalling them to zero, we thus delete in the equation for $\bar{v}$ the terms with too small coefficients.

Solving the equation for $\bar{v}$ obtained after the deletion we find approximate values of the roots having the greatest moduli. Then turning back to z we receive approximations to the roots of equation (30) with the greatest moduli. The greater m, the greater the accuracy of the approximations. There appears a difficulty in the transition from $\bar{v}$ to z connected with the extraction of a root with the index of radical 2^m. If equation (30) has real coefficients and if we get only one root with the greatest modulus in the equation for $\bar{v}$ then equation (30) will also have only one root with the greatest modulus and it will be real (why is it so?). Hence, in this case we can limit ourselves to calculating only two real values of the root. But if these conditions do not hold we have to extract the root according to the rules of Sec. 2 and thus obtain many possible values of the root. To determine which of the values should be taken we may substitute them all (in succession) into equation (30) and thus verify which of the roots satisfies the equation in the best way.

It is also possible to use the well-known relations between the roots of an algebraic equation and its coefficients. To deduce these relations it is necessary to compare the coefficients in equal powers of z in formulas (23) and (25) removing the parentheses in (25). This yields

$$\left.\begin{array}{c} z_1 + z_2 + \ldots + z_n = -\dfrac{a_1}{a_0} \\[2mm] z_1 z_2 + z_1 z_3 + \ldots + z_{n-1}z_n = \dfrac{a_2}{a_0} \\[2mm] \text{(all the possible products of two factors)} \\[2mm] \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\[2mm] z_1 z_2 \ldots z_n = (-1)^n \dfrac{a_n}{a_0} \end{array}\right\} \tag{33}$$

After the root v_1 of equation (31) has been found it is possible to eliminate the binomial $v - v_1$ from the equation by dividing

the equation by the binomial. Then in like manner we can find the next root etc. It is often possible to find the roots by applying relations (33) to equation (31). Doing this we should take into account that in case the roots of equation (31) differ greatly in the values of their moduli it is permissible to neglect certain summands in the left-hand sides of relations (33).

When the roots of equation (30) are determined approximately it is possible to apply some of the iterative methods described in § V.1, for example, Newton's method (Sec. V.2), to make the approximations more accurate.

Now let us consider as an example the equation $z^3 + z^2 - 3 = 0$ already solved in Sec. V.2. The successive transformations according to the Lobachevsky method yield

$$p^3 - p^2 + 6p - 9 = 0, \quad q^3 - 11q^2 + 18q - 81 = 0,$$
$$u^3 - 85u^2 + 2106u - 6561 = 0$$

(verify it!). Here we stop the process and make the substitution $u = 10^l \bar{u}$ which results in

$$\bar{u}^3 - \frac{85}{10^l}\bar{u}^2 + \frac{2106}{10^{2l}}\bar{u} - \frac{6561}{10^{3l}} = 0 \tag{34}$$

In this case we must take $l = 2$ (verify it by choosing another value of l). Then the absolute term in equation (34) becomes small and may be omitted. This implies

$$\bar{u}^2 - 0.85\bar{u} + 0.2106 = 0 \tag{35}$$

Hence, $\bar{u}_1$ and $\bar{v}_2$, and therefore z_1 and z_2, are pairwise conjugate imaginary numbers. Further, since $\bar{u}_1\bar{u}_2 = 0.2106$ we have $u_1u_2 = 2106$ and, by the equality $(z_1z_2)^8 = u_1u_2$, we deduce

$$z_1z_2 = |z_{1,2}|^2 = \sqrt[8]{2106} = 2.60$$

But $z_1z_2z_3 = 3$ (why is it so?), i.e. $z_3 = \dfrac{3}{2.60} = 1.15$.

Besides, $z_1 + z_2 + z_3 = -1$ (why is it so?). Therefore if we put $z_{1,2} = \alpha \pm i\beta$ then $2\alpha + 1.15 = -1$ which implies $\alpha = -1.08$. But $\alpha^2 + \beta^2 = |z_{1,2}|^2 = 2.60$, that is $\beta = \sqrt{2.60 - 1.08^2} = 1.20$. Thus, approximately, $z_{1,2} = -1.08 \pm i\,1.20$ and $z_3 = 1.15$.

Iterations by Newton's formula (V.6) yield more accurate values $z_{1,2} = -1.087 \pm i1.172$ and $z_3 = 1.175$.

The Lobachevsky method is especially convenient when the roots of the equation are real and different in their moduli. It is sometimes possible to simplify the calculations. For instance, in solving the above problem it was possible to limit the calculations to finding only z_3 because after dividing by $z - z_3$ we can reduce the problem to a quadratic equation.

COMPLEX NUMBERS AND FUNCTIONS 277 .

The Lobachevsky method and some similar methods are treated in detail in [22], [28], [38] and [50].

10. Decomposition of a Rational Fraction into Partial Rational Fractions. Remember that a rational fraction (see Sec. I.17) is a ratio of two polynomials:

$$f(z) = \frac{Q(z)}{P(z)} = \frac{b_0 z^m + \ldots + b_m}{a_0 z^n + \ldots + a_n} \tag{36}$$

If $m < n$ the fraction is called *proper* (no matter what the values of the coefficients are) and it is called *improper* if otherwise. An improper fraction can always be represented as a sum of an entire rational function (i.e. a polynomial) and a proper fraction. For instance, we can achieve this by dividing the numerator by the denominator according to the usual rule of division of polynomials. For example,

$$\frac{z^3}{z^2 - 3} = z + \frac{3z}{z^2 - 3}, \qquad \frac{z^2 - 1}{2z^2 + 5} = \frac{1}{2} + \frac{-\frac{7}{2}}{2z^2 + 5}$$

and so forth.

It is important to remark that a sum of proper rational fractions is also a proper rational fraction whereas this rule does not hold for fractional numbers. To prove this let us mark off the degree of a polynomial by the subscript. Then we have

$$\frac{Q_m(z)}{P_n(z)} + \frac{\overline{Q}_{\overline{m}}(z)}{\overline{P}_{\overline{n}}(z)} = \frac{Q_m(z)\,\overline{P}_{\overline{n}}(z) + \overline{Q}_{\overline{m}}(z)\,P_n(z)}{P_n(z)\,\overline{P}_{\overline{n}}(z)} \cdot$$

If the fractions on the left-hand side are proper then we have $m < n$ and $\overline{m} < \overline{n}$. But in this case the first summand in the numerator of the expression on the right-hand side is of degree $m + \overline{n} < n + \overline{n}$ and the second summand is of degree $\overline{m} + n < n + \overline{n}$. The whole numerator is therefore of degree $< n + \overline{n}$ whereas the degree of the denominator equals $n + \overline{n}$. Hence, the sum is a proper fraction. Obviously, the same is true for any number of summands.

Let (36) be a proper fraction. Factor the denominator in linear factors and combine similar factors [see formula (27)]. Then we can prove that the fraction can be represented in the form

$$\frac{Q(z)}{P(z)} = \frac{A_1}{(z - z_1)^{\alpha_1}} + \frac{A_2}{(z - z_1)^{\alpha_1 - 1}} + \ldots + \frac{A_{\alpha_1}}{z - z_1} + \frac{B_1}{(z - z_2)^{\alpha_2}} +$$

$$+ \frac{B_2}{(z - z_2)^{\alpha_2 - 1}} + \ldots + \frac{B_{\alpha_2}}{z - z_2} + \ldots + \frac{D_1}{(z - z_h)^{\alpha_h}} +$$

$$+ \frac{D_2}{(z - z_h)^{\alpha_h - 1}} + \ldots + \frac{D_{\alpha_h}}{z - z_h} \tag{37}$$

where A_1, A_2, ..., D_{α_k} are certain numerical coefficients. The rational fractions appearing on the right-hand side are called **partial rational fractions of the first type.** Thus, every proper rational fraction may be represented as a sum of partial rational fractions of the first type.

To prove the assertion we transform the fraction as follows:

$$\frac{Q(z)}{P(z)} = \frac{Q(z)}{a_0 (z-z_1)^{\alpha_1} (z-z_2)^{\alpha_2} \ldots (z-z_k)^{\alpha_k}} =$$

$$= \frac{Q(z)\,[(z-z_1)-(z-z_2)]}{a_0 (z-z_1)^{\alpha_1} (z-z_2)^{\alpha_2} \ldots (z-z_k)^{\alpha_k} (z_2-z_1)} =$$

$$= \frac{Q(z)}{a_0 (z_2-z_1) (z-z_1)^{\alpha_1-1} (z-z_2)^{\alpha_2} \ldots (z-z_k)^{\alpha_k}} -$$

$$- \frac{Q(z)}{a_0 (z_2-z_1) (z-z_1)^{\alpha_1} (z-z_2)^{\alpha_2-1} \ldots (z-z_k)^{\alpha_k}}$$

The constant factor $z_2 - z_1$ may be combined with the coefficient a_0 and thus the number of linear factors in the denominators of the two resulting fractions is by one less than that of the original fraction. Repeating transformations of this kind for each fraction we again reduce the number of linear factors in the denominators by one and so forth. We can proceed in this way as long as there are at least two different factors in the denominators. After a certain number of transformations there will no longer be different linear factors in the denominators, that is we shall arrive at a sum of fractions of the form $\dfrac{Q(z)}{a(z-z_l)^{\alpha}}$.

If we expand the numerator into powers of $z - z_l$ (see the beginning of Sec. IV.15) we obtain

$$\frac{Q(z)}{a(z-z_l)^{\alpha}} = \frac{c + d(z-z_l) + \ldots + g(z-z_l)^m}{a(z-z_l)^{\alpha}} = \frac{\dfrac{c}{a}}{(z-z_l)^{\alpha}} + \frac{\dfrac{d}{a}}{(z-z_l)^{\alpha-1}} + \cdots$$

Here after the division is performed there may be an entire part (that is an entire rational function, a polynomial). If now we add together all the fractions thus obtained we shall receive final formula (37), and all the entire parts must mutually cancel because otherwise a proper fraction would appear in the form of a sum of a proper fraction and a polynomial, which is impossible. (Why is it impossible?)

Practically, the decomposition is usually carried out by means of the *method of undetermined coefficients.* For this purpose we write the right-hand side of formula (37) with literal coefficients in the numerators and then find the coefficients. There are different

techniques for calculating the coefficients. We shall demonstrate them by considering an example.

Let it be required to decompose the fraction

$$\frac{x^3 - 2x + 3}{x\,(x-1)\,(x+2)^2} \tag{38}$$

into partial fractions. Since the fraction is proper we write, by formula (37),

$$\frac{x^3 - 2x + 3}{x\,(x-1)\,(x+2)^2} = \frac{A}{x} + \frac{B}{x-1} + \frac{C}{(x+2)^2} + \frac{D}{x+2} \tag{39}$$

where the coefficients, for simplicity's sake, are all denoted by different letters. Multiplying by the common denominator we receive

$$x^3 - 2x + 3 = A\,(x-1)\,(x+2)^2 + Bx\,(x+2)^2 +$$
$$+ Cx\,(x-1) + Dx\,(x-1)\,(x+2) \tag{40}$$

The equality must be an identity. We can therefore remove the parentheses and equate the coefficients of the same degrees of x. This yields the system of equations (verify it!) of the form

$$\begin{array}{r|l}
x^3 & A+B \qquad\;\; +D = 1 \\
x^2 & 3A+4B+C+D = 0 \\
x & 4B-C-2D = -2 \\
1 & -4A \qquad\qquad\;\; = 3
\end{array} \tag{41}$$

from which it is easy to find

$$A = -\frac{3}{4} = -0.750; \quad B = \frac{2}{9} = 0.222; \quad C = -\frac{1}{6} = -0.167;$$

$$D = \frac{55}{36} = 1.528 \tag{42}$$

This method is the most reliable [equating the coefficients we are sure that relation (40) is an identity and therefore (39) is also an identity] but not the simplest. There is another method which we can illustrate by taking (39) as an example. Without removing the parentheses in equality (40) we simply make x assume four different values according to the number of unknown coefficients to obtain equations for determining A, B, C and D. In our example it is very convenient to put $x = 0$, 1 and -2 since these values eliminate some summands and, in addition, to equate x to any arbitrary value, for example, to -1. This results in the relations

$$-4A = 3; \quad 9B = 2; \quad 6C = -1; \quad -2A - B +$$
$$+ 2C + 2B = 4 \tag{43}$$

which imply the same values (42).

The last method is the *collocation method*, that is the method of equating two expressions for different values of the argument. Such a method is especially effective in case the denominator of the fraction that should be decomposed has no multiple roots, that is all the linear fractions entering in its factorization have the first power.

If fraction (36) has real coefficients and if the independent variable is considered real the denominator can nevertheless have imaginary roots. Then though decomposition (37) is possible it may sometimes be inconvenient. In such cases another decomposition is often used. Namely, departing from decomposition (37) and using formula (29) we can prove the validity of the following decomposition:

$$\frac{Q(x)}{P(x)} = \frac{Q(x)}{a_0(x-x_1)^{\alpha_1} \ldots (x-x_r)^{\alpha_r}(x^2+p_1 x+q_1)^{\beta_1} \ldots (x^2+p_s x+q_s)^{\beta_s}} =$$

$$= \frac{A_1}{(x-x_1)^{\alpha_1}} + \frac{A_2}{(x-x_1)^{\alpha_1-1}} + \ldots + \frac{A_{\alpha_1}}{x-x_1} + \ldots + \frac{D_1}{(x-x_r)^{\alpha_r}} +$$

$$+ \frac{D_2}{(x-x_r)^{\alpha_r-1}} + \ldots + \frac{D_{\alpha_r}}{x-x_r} + \frac{M_1 x + N_1}{(x^2+p_1 x+q_1)^{\beta_1}} +$$

$$+ \frac{M_2 x + N_2}{(x^2+p_1 x+q_1)^{\beta_1-1}} + \ldots + \frac{M_{\beta_1} x + N_{\beta_1}}{x^2+p_1 x+q_1} + \ldots + \frac{U_1 x + V_1}{(x^2+p_s x+q_s)^{\beta_s}} +$$

$$+ \frac{U_2 x + V_2}{(x^2+p_s x+q_s)^{\beta_s-1}} + \ldots + \frac{U_{\beta_s} x + V_{\beta_s}}{x^2+p_s x+q_s} \tag{44}$$

The fractions on the right-hand side which have denominators equal to powers of quadratic trinomials are called the **partial rational fractions of the second type**. As before, all the coefficients entering into the numerators can be found by the method of undetermined coefficients. But since all the operations are performed now only on real numbers the unknown coefficients which should be found from a system of equations of the first degree [obtained by analogy with system (41) or (43)] must necessarily be real.

Thus, every proper rational fraction with real coefficients can be represented in the form of a sum of partial rational fractions of the first and of the second types with real coefficients. It may happen that there will be only fractions of the first type in case all the roots of the denominator are real or only fractions of the second type if all the roots are imaginary.

We shall not give here the proof of the possibility of decomposition (44). In every concrete problem the validity of (44) is confirmed by the results of the calculations.

11. Some General Remarks on Functions of a Complex Variable. Each of the functions

$$w = z^3 - iz, \quad w = \frac{z}{z^2+1}, \quad w = ze^z \text{ etc.} \tag{45}$$

assumes complex values for complex values of z. In the general form a relationship of this type can be written as $w = f(z)$ where both z and w are complex. Many notions of the theory of real variables can be transferred to the theory of complex functions of complex variables without any essential changes. This refers to the notion of a limit and to the properties of limits (of course, with the necessary exception of those properties that are connected with inequalities), to the notion of the continuity and to that of points of discontinuity and so on. The determination of the points of discontinuity of a function is carried out in the same way as for the functions of real variables in Sec. III.13. For instance, the first and the third of functions (45) are continuous for all z while the second one has two points of discontinuity $z = \pm i$ at which it approaches infinity.

The definition of the derivative of a function of a complex variable is analogous to the definition of the derivative of a real function (see Sec. IV.2):

$$\frac{dw}{dz} = f'(z) = \lim_{\Delta z \to 0} \frac{\Delta w}{\Delta z} = \lim_{\Delta z \to 0} \frac{f(z+\Delta z)-f(z)}{\Delta(z)}$$

where the limit must be uniquely determined and independent of the process of approaching zero as $\Delta z \to 0$ which can be quite arbitrary. One can easily verify that all the properties of the derivative
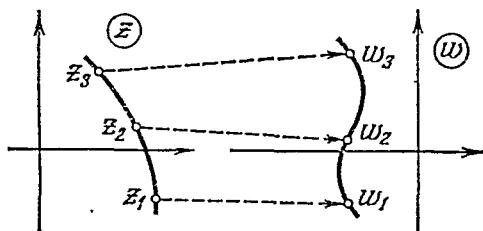
Fig. 188
The dotted lines represent the mapping: $w_1 = f(z_1)$ etc.

and all the differentiation formulas established in Secs. IV.4-5 remain true without changes but we are not going to treat these questions in detail here. The notion of derivatives of higher orders and the Taylor formula and series (see Sec. IV.5) that are based on this notion also remain true. We shall again discuss the question in Sec. XVII.14.

The geometric interpretation of a function of a complex variable $w = f(z)$ involves some essentially new ideas. Since the values of z are represented by points in a complex plane of the argument z and the corresponding values of w are represented by points in the complex plane of the variable w we see that in this case there is a certain correspondence between the points of the z-plane and the

points of the w-plane. We can also say that the points of the z-plane are *mapped* on the points of the w-plane. Or, in other words, the z-plane or its part where the function $w = f(z)$ is defined is mapped into the w-plane. For example, the function $w = z^3 - iz$ performs a mapping under which the points $z = 0$, $z = 1$, $z = i$ and $z = 2 - i$ are mapped on the points $w = 0$, $w = 1 - i$, $w = 1 - i$ and $w = 1 - 13i$, respectively, and so on. (Check it up!) If the point z moves in the z-plane and traces a curve then w also describes a certain curve in the w-plane (see Fig. 188). Thus, curves are transformed into curves under the mapping $w = f(z)$ and geometric figures in the z-plane are mapped on the geometric figures in the w-plane although the form of a geometric figure may be considerably changed by the mapping.

Functions
of Several Variables

§ 1. Functions of Two Variables

1. Methods of Representing. The concept of a function of any number of independent variables and the corresponding notation were introduced in Sec. I.11 and Sec. I.12. We have already used the concept but it is necessary to discuss the ways of representing such functions in greater detail.

The method of analytical representation of a function $z = f(x, y)$ depending on two variables does not essentially differ from the one applied to functions of one variable whereas the tabular method becomes much more complicated in this case because now we have to represent the values of two independent variables and it is therefore necessary to use a table with two entries. Such a table may have the following form:

TWO-ENTRY TABLE

$$z = f(x, y)$$

y / x	y_1	y_2	y_3	$\ldots$	y_N
x_1	$z_{11}=f(x_1, y_1)$	$z_{12}=f(x_1, y_2)$	$z_{13}=f(x_1, y_3)$	$\ldots$	$z_{1N}=f(x_1, y_N)$
x_2	$z_{21}=f(x_2, y_1)$	$z_{22}=f(x_2, y_2)$	$z_{23}=f(x_2, y_3)$	$\ldots$	$z_{2N}=f(x_2, y_N)$
$\cdot$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_M	$z_{M1}=f(x_M, y_1)$	$z_{M2}=f(x_M, y_2)$	$z_{M3}=f(x_M, y_3)$	$\ldots$	$z_{MN}=f(x_M, y_N)$

Here we have to denote the values of the function by two indices. Obviously, it is difficult to compile such a table if we have a great number of values for x and y.

"γραφειν" to write) which deals with methods of constructing *nomograms*, that is special drawings that are helpful for representing functions of any number of independent variables in a practically
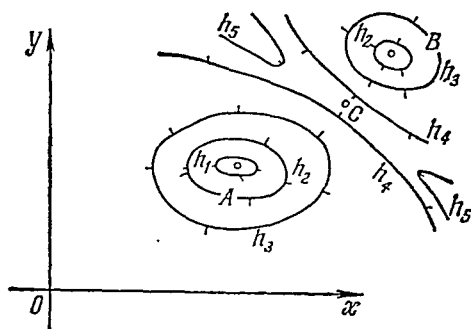
Fig. 192

convenient way. The application of nomograms saves much time and effort in calculations and does not require any special qualification. It should therefore be widely recommended. There are
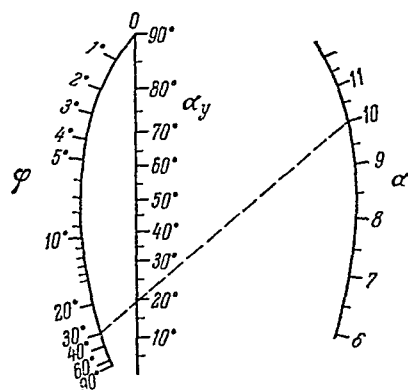
Fig. 193

many different types of nomograms. As an example we represent a nomogram in Fig. 193 which is designed for calculating one of the setting angles α_y of a cutter in a cutter grinder for given tool angles α and φ. The values of α, φ and α_y are marked on the three corresponding axes (two of them are curvilinear). If we apply a ruler to certain points α and φ on the corresponding axes we can read the desired value of α_y on the third axis. For instance, in Fig. 193 we have the values $\alpha = 10°$ and $\varphi = 30°$ which yield $\alpha_y = 19.5°$.

2. Domain of Definition. The domain of definition of a function $z = f(x, y)$ is the range of the independent variables x and y. If the independent variables are continuous the domain is either the whole x, y-plane or a region in the plane, or, finally, a totality of a number of regions in the plane. When we speak about a region in the x, y-plane we usually mean a *connected (simply-connected)* set of points, that is a set consisting of one entire part of the plane,

which does not degenerate (this means that it is not a line or a point or a number of separate points). We sometimes distinguish between *closed regions* (domains) which include their boundaries and *open* ones which do not include the boundaries. In other words, such regions in a plane play the same role as intervals on a straight line (see Sec. I.5).

For example, the domain of definition of the function $z = x + y$ is the whole x, y-plane. The domain of the function $z = \sqrt{y - x}$
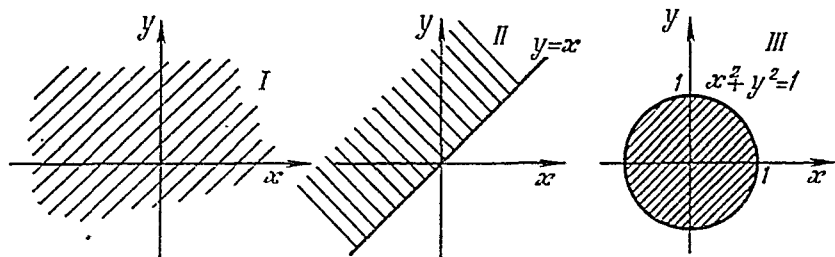
Fig. 194

(in case we consider only real values of z) is defined by the inequality $y - x \geqslant 0$, i.e. $y \geqslant x$. The domain of the function $z = \dfrac{1}{\sqrt{1 - x^2 - y^2}}$ is obtained from the inequality $x^2 + y^2 < 1$ etc. These domains are shown in Fig. 194.

3. Linear Function. According to Sec. I.17, a linear function of two variables has the form

$$z = ax + by + c \tag{1}$$

where a, b and c are constant coefficients. By analogy with Sec. I.22 we can easily derive a formula for an increment of the function:

$$\Delta z = a\Delta x + b\Delta y$$

Similar formulas hold for linear functions of any number of variables.

Formula (1) having three coefficients, any **linear approximation** (i.e. an approximate replacement of a function by a linear function) requires three conditions. For instance, let the values of a function $f(x, y)$ be known:

$$f(x_1, y_1) = z_1, \quad f(x_2, y_2) = z_2, \quad f(x_3, y_3) = z_3$$

If we want to construct a linear function (1) which takes on the same values (i.e. to carry out the **linear interpolation**) by substi-

tuting (approximately) an expression of type (1) for f we must have

$$\left.\begin{array}{l} ax_1 + by_1 + c = z_1 \\ ax_2 + by_2 + c = z_2 \\ ax_3 + by_3 + c = z_3 \end{array}\right\} \tag{2}$$

We determine the coefficients by solving this system. Such a replacement of f by (1) yields good results if, in the first place, we consider the values of the arguments lying inside the triangle with the vertices (x_1, y_1), (x_2, y_2) and (x_3, y_3) (see Fig. 195) and, in the second place, if the triangle is not large enough for non-linear properties of the function f to be manifested in a noticeable manner. Besides, the triangle must not have very acute angles. (If in a certain limiting process one of the angles vanishes and the triangle turns into a line segment the determinant of system (2) also vanishes and our calculations become inapplicable.)
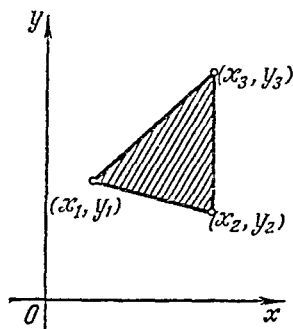
Fig. 195

If we replace f by (1) outside the triangle (this is the linear extrapolation) then, in general, the error will increase as we move away from the triangle.

We can likewise carry out linear interpolations for functions of any number of arguments.

4. Continuity and Discontinuity. The concept of continuity of a function $z = f(x, y)$ is quite similar to that of a function of one argument which was discussed in Sec. I.16 and Sec. III.12. As an example we can formulate the following definition of continuity: a function f is called a continuous function for the values $x = x_0$ and $y = y_0$ of the arguments if for every process in which $x \to x_0$ and $y \to y_0$ (in an arbitrary way) we have $f(x, y) \to f(x_0, y_0)$. If otherwise the function is said to be discontinuous for these values of the arguments. Then the point with the coordinates (x_0, y_0) lying in the x, y-plane is called the point of discontinuity of the function. A function which is continuous at each point of a region is said to be continuous in the region.

It should be noted that besides separate points of discontinuity a function of two variables may have entire *lines of discontinuity*, that is lines wholly consisting of points of discontinuity. For instance, let us take the functions

$$z = \frac{1}{x^2 + y^2} \quad \text{and} \quad z = \frac{1}{(y - x)^2}$$

The first function has only one point of discontinuity $(0, 0)$ whereas the second function has an entire line of discontinuity, namely, the straight line $(y - x)^2 = 0$, i.e. $y = x$. The level lines of the functions are depicted in Fig. 196. In both cases the functions approach infinity at the points of discontinuity. But, as in the case of a function of one variable, there are other types of discontinuities. In practical problems we often encounter such a line of discontinuity of a function that in approaching any point of this line from one side the function has a certain finite limit whereas in approaching
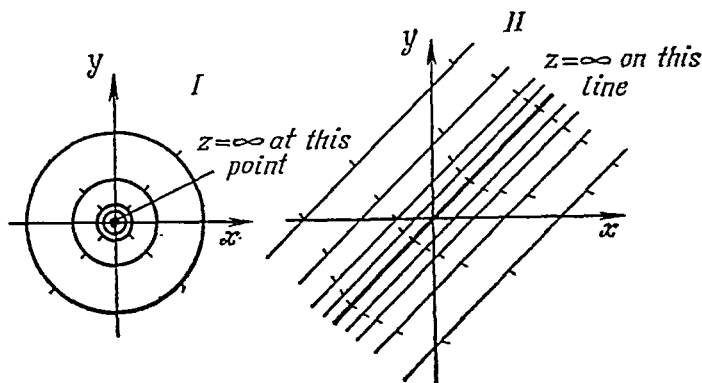
Fig. 196

the same point from the other side of the line the function has a different finite limit. In such a case the function has a finite *jump* as the point (x, y) passes through the line. An approximate sketch of the graph of such a function is depicted in Fig. 197.

The behaviour of a function of two variables in the vicinity of its point of discontinuity may essentially depend on the way the point is approached. For instance, there may exist a limit depending on the choice of a path of approaching the point for one group of paths and there may exist neither a finite nor an infinite limit for other paths etc. Since there is an infinitude of ways of approaching a point of discontinuity (whereas we have only two main ways of approaching a point of discontinuity in the case of a function of one argument, namely, from the right or from the left side) points of discontinuity of functions of several variables are, in general, of a more complicated type than those of functions of one independent variable. For example, the function $z = \dfrac{2xy}{x^2+y^2}$ has its only point of discontinuity at the point $x = 0$, $y = 0$ where the denominator vanishes. Now if $x \to 0$ and $y \to 0$ in such a way that $\dfrac{y}{x} = k$, i.e. $y = kx$, where k is a constant we obtain $z = \dfrac{2xkx}{x^2+k^2x^2} = \dfrac{2k}{1+k^2}$

and therefore the limit depends on the relation between y and x (see Fig. 198). If in approaching a point there is no single finite or infinite limit of a function then calculating the limit at the point we must specify the way of approaching it to avoid misunderstandings.

The properties of continuous functions of two arguments defined in a finite closed region in the x, y-plane are analogous to those described in Sec. III.14 for functions of one argument defined over a closed finite interval. We are therefore not going to enumerate the properties again.
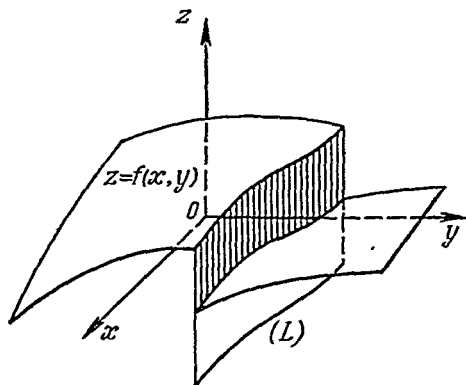
Fig. 197

L) is the line of discontinuity of the function f; it lies in the plane xOy
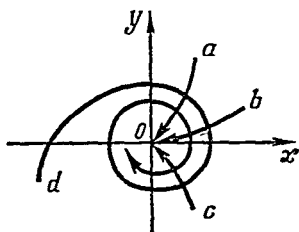
Fig. 198

In approaching the origin in the direction a the limit of z is equal to 1; the limit is equal to zero in the direction b and it is equal to -1 in the direction c. There is no limit in approaching the origin along the spiral d

We sometimes encounter the problem of solving an inequality of the form $f(x, y) > 0$. This can be carried out in a way similar to the method of solving inequalities discussed in Sec. III.15 [see inequality (III.17)]. We should first draw the curve $f(x, y) = 0$ and the lines of discontinuity of the function f provided there are such lines. All these curves break the plane into parts, and the function retains its sign inside each of the parts. We can determine the corresponding signs by calculating the values of the function for the points arbitrarily chosen in each of the parts.

For instance, let us solve the inequality

$$\frac{x^2+y^2-4}{x+y} > 0 \qquad (3)$$

Here the circle $x^2 + y^2 - 4 = 0$ is the *line of zeros* and the straight line $x + y = 0$ serves as the line of discontinuity. They divide the plane into four parts shown in Fig. 199. Now we take a point in each of the parts, for example, the points $(-3, 0)$, $(-1, 0)$,

(1, 0) and (3, 0), and determine the corresponding signs of the function which are —, +, — and +. The regions where inequality (3) holds are shaded in Fig. 199.

5. Implicit Functions. The definition of an implicit function of two arguments is similar to that in Sec. I.20 given for a function of one independent variable. An implicit function $z\,(x, y)$ is defined by an equation of the form

$$F\,(x, y, z) = 0 \qquad (4)$$

Here, as in Sec. I.20, the function $z\,(x, y)$ may turn out to be multiple-valued and then we can consider its *single-valued branches*.

Equation (4) may define a surface of an arbitrary form whereas

Fig. 199

a surface determined by an equation $z = f\,(x, y)$ is punctured by any straight line parallel to the z-axis at not more than one point (see Fig. 190).
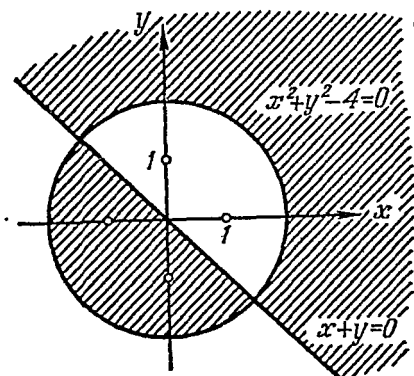
§ 2. Functions of Arbitrary Number of Variables

6. Methods of Representing. The main notions related to the analytical form of a function and to its properties are transferred to functions of any number of arguments. But there are some additional difficulties in investigating such functions. First of all, the tabular and the graphical ways of their representation become too complicated. We can, of course, represent a function of three variables by means of a system of two-entry tables (see Sec. 1) or by a set of pictures similar to Figs. 189 or 192, but this is very difficult.

But in certain cases the calculation of the values of a function of a large number of arguments may sometimes be reduced to the calculation of the values of several functions of a lower number of variables. Then we can widely use the methods described in Sec. I.13 and Sec. 1. For instance, take a function of four arguments of the form $u = f\,(x, y) + \varphi\,(z, t)$. To calculate the values of u we need tables and graphs of the functions f and φ. But each of the last two functions depends only upon two variables and this facilitates the calculations. Similarly, the calculation of the values of the function $u = f\,[\varphi\,(x) + y,\ \psi\,(z) - t]$ depending on four independent variables requires the representation of one function of two variables and of two functions of one variable and so on. In such cases the calculation and the investigation of functions become much easier. Unfortunately, not all functions can be represented in this way.

7. Functions of Three Arguments. Another difficulty lies in the geometrical interpretation of a functional relationship in case the whole number of variables (both dependent and independent) is greater than 3, that is greater than the dimension of our usual geometrical space. The situation is comparatively simpler for a function of three arguments $u = f(x, y, z)$ (the total number of variables equals 4 here). In this case the domain of definition is either the whole space of the variables x, y and z or some part of it, that is one or several regions (domains) in the x, y, z-space (see Sec. 2; the notion of a degenerated region should be appropriately changed in this case). We can therefore represent such a domain geometrically. For instance, the domain of definition of the function $u = x^2 + y^2 - z$ is the whole space whereas the function $u = \sqrt{1 - x^2 - y^2 - z^2}$ is defined only if

$$1 - x^2 - y^2 - z^2 \geqslant 0 \quad \text{or} \quad x^2 + y^2 + z^2 \leqslant 1$$

i.e. in the last case the domain is the sphere of radius 1 with centre at the origin of coordinates.

We can also consider *level surfaces* of a function $f(x, y, z)$ understanding them in the sense of Sec. 1, that is as surfaces in the x, y, z-space on which the function is constant: $f(x, y, z) = \text{const}$.

Points of discontinuity (provided a function has them) lie in the x, y, z-space and can also be represented in a visual way. The points of discontinuity of a function of three independent variables can be located separately but they can also form lines of discontinuity and even *surfaces of discontinuity*, that is surfaces which entirely consist of points of discontinuity. For example, when investigating a passage from one physical medium into another we interpret the interface as a surface on which the quantities characterizing the properties of the media have discontinuities (for instance, this is applicable to investigating the passage of the light from water into air or from glass into air etc.).

8. General Case. The concept of a space of variables is quite visual and it is therefore convenient to transfer it to the case of functions of any number of independent variables. This leads to the notion of a many-dimensional Cartesian space (see Sec. VII.18, example 3). For instance, let a function $w = f(x, y, z, u)$ of four arguments be considered. Then every quadruple of values x, y, z and u determines a point in the space E_4 (strictly speaking, such a quadruple is, by definition, a point of E_4). Thus the space E_4 serves as the space of arguments here; the domain of definition of the function w is a region (domain) in the space E_4 or a set of a number of such regions. The function may be continuous or may have separate points of discontinuity, lines of discontinuity, two-dimensional surfaces or three-dimensional *hypersurfaces* (see Sec. VII.19) consisting of points of discontinuity.

To construct the "graph" of a function $u = f(x, y, z, t)$ we need the five-dimensional space E_5 of the variables x, y, z, t and u. To find the points of the graph we must make x, y, z and t assume arbitrary values and calculate the corresponding values of u. [For instance, we can easily verify that the graph of the function $u = xz - 2y^2 t$ passes through the points $(1, 1, 2, 0, 2)$ and $(-1, 2, 0, -2, 16)$ etc.] When we discussed functions of three variables in Sec. 7 we saw that the space of the arguments had a visual geometrical interpretation. But this is not so for its graph that lies in the four-dimensional space E_4 which we cannot visualize.

We shall see in Sec. X.2 that a many-dimensional space can be interpreted in a less formal way than that connected with the notion of a Cartesian space of n-tuples consisting of n numbers.

9. Concept of Field. We say that there is a **field** of a quantity in space if a certain value of the quantity is defined at each point of the space. For instance, when investigating a flow of gas we consider the field of temperature (the temperature has a certain value at each point), the field of density, the field of velocities and so on. A field can be a **scalar field** or a **vector field** depending on the properties of the quantity in question. For example, a temperature field and a density field are scalar ones whereas a velocity field or a field of force is a vector field. A field is called **stationary (steady-state)** if it does not change at each point of the space as time passes and it is called **non-stationary** if such a change takes place.

For definiteness, let us denote a scalar quantity by the letter u and an arbitrary (variable) point in space by the letter M. Then to each position of the point M there corresponds a certain value of the quantity u and we can therefore regard u as a function of M: $u = f(M)$. Such a function differs from those considered above since a point is not a quantity. But in the general sense of the notion of a function (widely used in modern mathematics) we can apply the term "function" to every situation when there exists a certain law according to which to the objects of one "kind" (of an arbitrary nature) there correspond the objects of some other "kind" (in our case the objects of the first kind are points in space and the objects of the second kind which correspond to the points are the values of the quantity u). In case a field is non-stationary we have $u = f(M, t)$ where t is the time.

Now we can easily pass from a *function of a point* to a function of three variables, namely, of three spatial coordinates. To do this it is sufficient to introduce a Cartesian coordinate system x, y, z in space. Then the position of a point M in space is completely characterized by the corresponding values x, y and z, that is we can write $u = u(x, y, z)$. Conversely, if a coordinate system x, y, z is given then any function of x, y and z can be regarded as a function of a point. But we should take into account that a field $u = f(M)$

(i.e. a function u of the point M) has its own meaning and can be investigated without introducing any coordinate system. Besides, introducing coordinate systems in different ways we obtain different relationships of the form $u = u\,(x,\,y,\,z)$ for one and the same function $u = f\,(M)$. Hence, when investigating a field we regard the concept of a function of a point as a primary concept relative to the concept of a function of the coordinates of a point.

If a quantity, according to its physical or geometrical meaning, depends on the position of a point in a plane we call the corresponding field a **plane field**. We encounter such fields when investigating thermal processes in a thin plate whose thickness can be neglected.

If we have a space field of a quantity u which depends only on x and y and does not depend on z in a certain coordinate system $x,\,y,\,z$ the field is said to be a **plane-parallel field**. Then we can regard such a field as being defined in the $x,\,y$-plane (and as independent of z) which means that the field can be treated as a plane field. But of course we should keep in mind that in reality the field is spatial and that the relationship between u and the coordinates is the same in all the planes parallel to the $x,\,y$-plane.

§ 3. Partial Derivatives and Differentials of the First Order

10. Basic Definitions. Let a function of several independent variables be given. For definiteness, let it be a function of three arguments $u = f\,(x,\,y,\,z)$. If we fix certain values of all the arguments but one the variable u becomes a function of this single argument. We can therefore differentiate the function with respect to the argument and take its differential as it was done in Chapter IV for a function of one independent variable. Such derivatives are called **partial derivatives**. The corresponding differentials are **partial differentials** of the function. In other words,

$$u_x' = f_x'\,(x,\,y,\,z) = \lim_{\Delta x \to 0} \frac{\Delta_x u}{\Delta x}\,,$$

$$u_y' = f_y'\,(x,\,y,\,z) = \lim_{\Delta y \to 0} \frac{\Delta_y u}{\Delta y}$$

where $\Delta_x u = f\,(x + \Delta x,\,y,\,z) - f\,(x,\,y,\,z)$ and $\Delta_y u = f\,(x,\,y + \Delta y,\,z) - f\,(x,\,y,\,z)$ are the **partial increments** of the function. A partial increment corresponds to a change of one of the variables when all the other variables are kept constant. Let the reader write the expression of the increment $\Delta_z u$ and of the derivative u_z'. One should take into account that the symbol u' or $f'\,(x,\,y,\,z)$ has no

sense for a function of several variables because one must necessarily indicate the variable with respect to which the derivative is taken.

The computation of partial derivatives of concrete elementary functions is performed according to the rules given in Sec. IV.5. When we differentiate with respect to a certain variable we must regard all the other variables as constants.

Example. Let $u = x^2z^3 - y^z$ then $u'_x = 2xz^3$, $u'_y = -zy^{z-1}$ and $u'_z = 3x^2z^2 - y^z \ln y$ (check up these results!).

A partial differential is denoted by the symbol ∂ with a subscript indicating the argument with respect to which the differentiation is performed. Thus,

$$\partial_x u = u'_x \, \Delta x, \quad \partial_y u = u'_y \, \Delta y \quad \text{and} \quad \partial_z u = u'_z \, \Delta z$$

In particular, if we put $u = x$ here we obtain

$$\partial_x x = \Delta x \quad \text{and} \quad \partial_y x = \partial_z x = 0 \tag{5}$$

since $x'_x = 1$ and $x'_y = x'_z = 0$. We have therefore $\partial_x u = u'_x \, \partial_x x$ which implies $u'_x = \frac{\partial_x u}{\partial_x x}$. One usually omits the subscripts in putting down the last formula and simply writes $u'_x = \frac{\partial u}{\partial x}$ because the denominator itself indicates that the derivative is taken with respect to x (or with respect to some other variable in other cases). Similarly, $u'_y = \frac{\partial u}{\partial y}$ and $u'_z = \frac{\partial u}{\partial z}$. This rule of writing differentials sometimes becomes inconvenient. Indeed, in the first place, not only denominators differ in the last two expressions but the numerators as well. Actually, the numerator of the first fraction must be regarded as $\partial_y u$ whereas the numerator of the second fraction as $\partial_z u$. In the second place, for example, we cannot write $\frac{\partial x}{\partial u}$ instead of $\left(\frac{\partial u}{\partial x} \right)^{-1}$ since the differentials in these expressions have a different sense (in the first expression the differentials are taken with respect to u whereas in the second one with respect to x).

When using partial derivatives one must be careful and pay much attention to the choice of independent variables. For instance, if we write the expression for the power of an electric current in the form $P = \frac{U^2}{R}$ where U is the voltage and R is the resistance of the circuit we obtain $\frac{\partial P}{\partial R} = -\frac{U^2}{R^2} = -\frac{P}{R}$. But if the same formula is written in the form $P = I^2 R$ where I is the flow of the electric current then we get $\frac{\partial P}{\partial R} = I^2 = \frac{P}{R}$. There is no contradiction between these results. Actually, if we write them in full we shall have $\frac{\partial P}{\partial R}\Big|_{U=\text{const}} = -\frac{P}{R}$ for the first result and $\frac{\partial P}{\partial R}\Big|_{I=\text{const}} = \frac{P}{R}$ for the

second one. (Let the reader think about the physical meaning of the signs $+$ and $-$ entering into the last formulas.)

11. Total Differential. The *total* (or *exact*) *differential* of a function $u = f(x, y, z)$ is equal to the sum of all its partial differentials:

$$du = \partial_x u + \partial_y u + \partial_z u = u'_x \Delta x + u'_y \Delta y + u'_z \Delta z =$$
$$= \frac{\partial u}{\partial x} \Delta x + \frac{\partial u}{\partial y} \Delta y + \frac{\partial u}{\partial z} \Delta z \qquad (6)$$

Formula (6) must be regarded as the definition of a total differential. In particular, if we put $u = x$ formula (5) implies that

$$dx = \partial_x x + \partial_y x + \partial_z x = \Delta x$$

and hence the total differential of an independent variable is equal to the increment of the variable (compare with Sec. IV.8). Formula (6) can therefore be rewritten as

$$du = u'_x dx + u'_y dy + u'_z dz = \frac{\partial u}{\partial x} dx + \frac{\partial u}{\partial y} dy + \frac{\partial u}{\partial z} dz \qquad (7)$$

For example,

$$d \left(x^2 \sin \frac{x}{y} \right) = \left(2x \sin \frac{x}{y} + \frac{x^2}{y} \cos \frac{x}{y} \right) dx - \frac{x^3}{y^2} \cos \frac{x}{y} dy$$

The connection between the total differential of a function and the total increment of the function is analogous to the one described in Sec. IV.8. Let the independent variables receive increments Δx, Δy and Δz. Then u receives the increment

$$\Delta u = f(x + \Delta x, y + \Delta y, z + \Delta z) - f(x, y, z)$$

This is the **total increment**. It can be represented in the form of a sum of three increments so that each increment corresponds to the change of one of the variables. Namely,

$$\Delta u = [f(x + \Delta x, y, z) - f(x, y, z)] +$$
$$+ [f(x + \Delta x, y + \Delta y, z) - f(x + \Delta x, y, z)] +$$
$$+ [f(x + \Delta x, y + \Delta y, z + \Delta z) - f(x + \Delta x, y + \Delta y, z)] \qquad (8)$$

The reader must carefully check up this formula! The increments in the square brackets entering into formula (8) are nothing but partial increments and are therefore connected with the partial differentials in the way described by formula (IV.23). Hence,

$$f(x + \Delta x, y, z) - f(x, y, z) = f'_x(x, y, z) \Delta x + \alpha_1 \Delta x,$$
$$f(x + \Delta x, y + \Delta y, z) - f(x + \Delta x, y, z) =$$
$$= f'_y(x + \Delta x, y, z) \Delta y + \alpha_2 \Delta y =$$
$$= [f'_y(x, y, z) + \alpha_3] \Delta y + \alpha_2 \Delta y =$$
$$= f'_y(x, y, z) \Delta y + \alpha_4 \Delta y,$$
$$f(x + \Delta x, y + \Delta y, z + \Delta z) - f(x + \Delta x, y + \Delta y, z) =$$
$$= f'_z(x, y, z) \Delta z + \alpha_5 \Delta z$$

(the last formula is deduced in like manner) where all the quantities denoted by α tend to zero as $\Delta x \to 0$, $\Delta y \to 0$ and $\Delta z \to 0$. Substituting these expressions into (8) we obtain

$$\Delta u = u'_x \, \Delta x + u'_y \, \Delta y + u'_z \, \Delta z + \alpha \, \Delta x + \beta \, \Delta y + \gamma \, \Delta z =$$
$$= du + \alpha \, \Delta x + \beta \, \Delta y + \gamma \, \Delta z \qquad (9)$$

where α, β and $\gamma \to 0$ as Δx, Δy and $\Delta z \to 0$. Consequently, in the general case we can also say that *the total differential is the principal linear part of the increment of a function.* We call it linear because it equals the sum of summands which are directly proportional
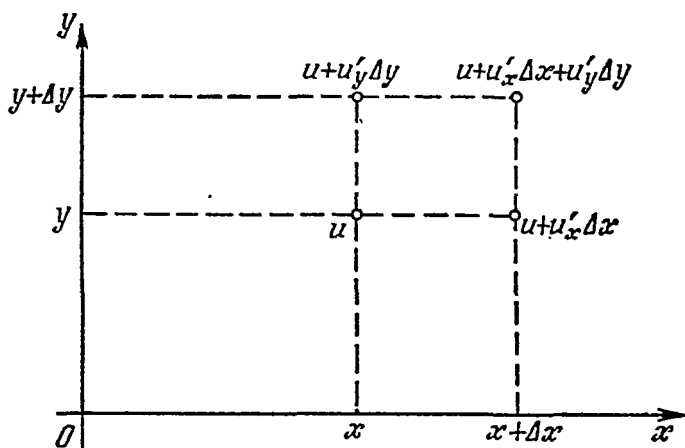
Fig. 200

to the increments of the arguments and it is regarded as the principal part of the total increment since it differs from the increment by an infinitesimal variable of higher order with respect to the increments of the arguments (compare with Sec. 3). As in the case of a function of one independent variable, the replacement of an increment of a function by its differential is equivalent to the replacement of a non-linear function by a linear one.

Formula (9) is illustrated in Fig. 200 for the case of two independent variables. The values of the function (with the infinitesimals of higher order neglected) are put down near the corresponding points of the x, y-plane.

As in Sec. IV.10, the approximate relations $\Delta_x u \approx \partial_x u$, $\Delta_y u \approx \partial_y u$, $\Delta_z u \approx \partial_z u$ and, particularly, $\Delta u \approx du$ are the source of many useful concrete approximate formulas. We note here that the last formula can be rewritten in full as

$$f(a + h, \ b + k, \ c + l) \approx f(a, \ b, \ c) + f'_x(a, \ b, \ c) \, h +$$
$$+ f'_y(a, \ b, \ c) \, k + f'_z(a, \ b, \ c) \, l$$

It is also easy to deduce the formula

$$\alpha_u = |f'_x(\overline{x},\ \overline{y},\ \overline{z})|\,\alpha_x + |f'_y(\overline{x},\ \overline{y},\ \overline{z})|\,\alpha_y + |f'_z(\overline{x},\ \overline{y},\ \overline{z})|\,\alpha_z$$

which is analogous to formula (IV.29) and is obtained in a similar way.

For instance, let $u = xy$. Then

$$\alpha_u = |\overline{y}|\,\alpha_x + |\overline{x}|\,\alpha_y,$$

$$\delta_u = \frac{\alpha_u}{|\overline{u}|} = \frac{|\overline{y}|\,\alpha_x + |\overline{x}|\,\alpha_y}{|\overline{xy}|} = \frac{\alpha_x}{|\overline{x}|} + \frac{\alpha_y}{|\overline{y}|} = \delta_x + \delta_y$$

that is the operation of multiplication (and, similarly, the operation of division) of approximate numbers yields the addition of their maximum relative errors (let the reader check up this rule for the case of division!).

12. Derivative of Composite Function. Let again $u = f(x,\ y,\ z)$ and let the variables no longer be independent but in their turn depend on some independent variables s and t. Thus we have

$$u = u(x,\ y,\ z), \quad x = x(s,\ t), \quad y = y(s,\ t), \quad z = z(s,\ t) \quad (10)$$

We see that u becomes a composite function of s and t. To calculate the partial derivative u'_s let us fix t and make s receive an increment Δs. Then x, y and z will also receive certain partial increments and therefore u also gains an increment which, according to formula (9), can be written in the form

$$\Delta_s u = u'_x\,\Delta_s x + u'_y\,\Delta_s y + u'_z\,\Delta_s z + \alpha\,\Delta_s x + \beta\,\Delta_s y + \gamma\,\Delta_s z$$

If now we divide both sides by Δs and pass to the limit as $\Delta s \to 0$ we shall obtain

$$u'_s = \frac{\partial u}{\partial s} = u'_x x'_s + u'_y y'_s + u'_z z'_s = \frac{\partial u}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial s} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial s} \quad (11)$$

The derivative u'_t is expressed similarly. Thus we see that the rule we have deduced is analogous to that of differentiating a function of one independent variable [see formula (IV.9)] but the number of summands is greater here since the derivatives with respect to all intermediate variables also enter into the differentiation formula.

Formula (11) implies [this is analogous to the results obtained in Sec. IV.9] that formula (7) [but not formula (6)!] remains true even in the case when the former independent variables turn out to be dependent on some other variables. In fact, in the case of formula (10) we have

$$du = u'_s\,ds + u'_t\,dt = (u'_x x'_s + u'_y y'_s + u'_z z'_s)\,ds +$$
$$+ (u'_x x'_t + u'_y y'_t + u'_z z'_t)\,dt = u'_x(x'_s\,ds + x'_t\,dt) +$$
$$+ u'_y(y'_s\,ds + y'_t\,dt) + u'_z(z'_s\,ds + z'_t\,dt) = u'_x\,dx + u'_y\,dy + u'_z\,dz$$

which is what we set out to prove. Hence, formula (7) is invariant, that is holds for all cases [just as it is for formula (IV.22)].

The *invariance* of the form of the total differential of a function implies many differentiation formulas. For instance, if $w = uv$ where u and v may depend on some other variables then formula (7) yields $dw = w'_u\, du + w'_v\, dv = v\, du + u\, dv$. Consequently, the formula $d\,(uv) = v\, du + u\, dv$ holds for all cases. In a similar way we can verify the validity of the following formulas:

$$d\,(u \pm v) = du \pm dv, \quad d\,(Cu) = C du \quad (C = \text{const}),$$

$$d\left(\frac{u}{v}\right) = \frac{v\, du - u\, dv}{v^2}, \quad d\,(u^n) = nu^{n-1}\, du, \quad d\,(\sin u) = \cos u\, du \text{ etc.}$$

In many cases these formulas enable us to calculate a total differential directly, without computing the corresponding partial derivatives.

For example, $d\,(\sin x^2 y^3) = \cos x^2 y^3 d\,(x^2 y^3) = \cos\,(x^2 y^3)\,[y^3 d\,(x^2) + x^2 d\,(y^3)] = \cos x^2 y^3\,(2xy^3\, dx + 3x^2 y^2\, dy)$. Conversely, the total differential of a function being given, we can restore the partial derivatives determining them by taking the coefficients in dx and dy.

Here we give several examples on calculating derivatives.

(1) Let $u = f\left(\sqrt{x^2 + y^2}\right)$. Then

$$u'_x = f'\left(\sqrt{x^2 + y^2}\right)\left(\sqrt{x^2 + y^2}\right)'_x = f'\left(\sqrt{x^2 + y^2}\right)\frac{2x}{2\sqrt{x^2 + y^2}} =$$

$$= \frac{x}{\sqrt{x^2 + y^2}}\, f'\left(\sqrt{x^2 + y^2}\right)$$

Here the function f itself is a function of one variable. This variable is replaced by $\sqrt{x^2 + y^2}$; the symbol f' designates the derivative of f with respect to its single argument.

(2) Let $u = f\left(\frac{x}{y}, \frac{y}{x}, y\right)$. Then

$$u'_y = f'_I\left(\frac{x}{y}, \frac{y}{x}, y\right)\left(-\frac{x}{y^2}\right) + f'_{II}\left(\frac{x}{y}, \frac{y}{x}, y\right)\frac{1}{x} + f'_{III}\left(\frac{x}{y}, \frac{y}{x}, y\right)$$

Here the function f itself is a function of three variables. The expressions $\frac{x}{y}$, $\frac{y}{x}$ and y are substituted, respectively, for these independent variables; the expressions f'_I, f'_{II} and f'_{III} designate the derivatives of f taken with respect to these three variables.

(3) Let $y = x^{\sin x}$. Then to compute y' we can take logarithms as it was recommended in the end of Sec. IV.5. But we can also use another method based on the above results. Let us denote $y = u^{\sin v}$ where $u = x$ and $v = x$ (such intermediate

variables are usually introduced mentally without putting them down). Now we can differentiate y as a composite function (u and v are regarded as intermediate variables):

$$y'_x = y'_u u'_x + y'_v v'_x = \sin v\, u^{\sin v - 1} \cdot 1 + u^{\sin v} \ln u \cos v \cdot 1 =$$
$$= \sin x\, x^{\sin x - 1} + x^{\sin x} \ln x \cos x$$

It is obvious that the last method is more general. If we have to compute the derivative with respect to x of an expression into which x enters several times we should differentiate with respect to each ("imagined") argument that involves x, multiply by the derivative with respect to x and add together the results obtained.

Here we are only going to consider functions of three variables but the following definition will also be useful for our further aims since it can be easily transferred to the case of functions of any number of independent variables. A function $F(x, y, z)$ is called a **homogeneous function of degree** k if for any $t > 0$ we have

$$F(tx, ty, tz) \equiv t^k F(x, y, z) \tag{12}$$

For instance, the function $F(x, y, z) = x^2 - 3yz$ is a homogeneous function of degree 2 because

$$F(tx, ty, tz) = (tx)^2 - 3\,(ty)\,(tz) = t^2\,(x^2 - 3yz) =$$
$$= t^2 F(x, y, z)$$

We can similarly verify that the function $\dfrac{x \sin\left(\dfrac{y}{z}\right)}{y - z}$ is a homogeneous function of degree zero, the function $\dfrac{1}{\sqrt{x - y - z}}$ is a homogeneous function of degree $-\dfrac{1}{2}$ whereas, for example, the function $x + 2y - z + 1$ is not a homogeneous one at all.

In the general case formula (12) can be written as $F(ta, tb, tc) = t^k F(a, b, c)$ for any a, b and c. If we differentiate with respect to t (and regard a, b and c as constants) we obtain

$$F'_x(ta, tb, tc)\, a + F'_y(ta, tb, tc)\, b + F'_z(ta, tb, tc)\, c =$$
$$= kt^{k-1} F(a, b, c)$$

Now putting $t = 1$, $a = x$, $b = y$ and $c = z$ in the last formula we deduce the formula

$$x F'_x(x, y, z) + y F'_y(x, y, z) + z F'_z(x, y, z) = k F(x, y, z)$$

which expresses so-called **Euler's theorem on homogeneous functions.**

13. Derivative of Implicit Function. Let an implicit function $z = z(x, y)$ be defined by an equation

$$F(x, y, z) = 0 \tag{13}$$

To calculate the partial derivative z'_x we must fix y and differentiate formula (13) with respect to x taking into account that z also depends on x. Thus, applying the rule of differentiating a composite function we obtain

$$F'_x x'_x + F'_z z'_x = 0, \quad \text{i.e.} \quad F'_x + F'_z z'_x = 0 \qquad (14)$$

which implies

$$z'_x = - \frac{F'_x(x, y, z)}{F'_z(x, y, z)} \qquad (15)$$

Similarly, $z'_y = - \dfrac{F'_y}{F'_z}$. To guarantee that this derivative assumes a finite value we should additionally introduce the requirement

$$F'_z(x, y, z) \neq 0 \qquad (16)$$

Formula (16) expresses a sufficient condition for the existence of an implicit function $z = z(x, y)$ defined by equation (13); the geometrical meaning of the condition will be illustrated in Sec. XII.3.

Implicit functions may be defined by a system of equations. Suppose we have m equations which are **compatible** (that is they do not contradict each other and can be solved, at least theoretically), **independent** (this means that none of the equations is the consequence of the others) and connect n variables. Then if $m < n$ (i.e. the number of equations is less than the number of independent variables) we can regard certain $n - m$ variables as independent. We can make them take on arbitrary values and express the remaining m variables as functions of these independent variables by solving the equations. (If the number of equations is equal or exceeds the number of variables then, generally, we get some discrete values and it is therefore impossible to construct functions.) As an example let us take the case of two equations containing five variables, namely,

$$\left.\begin{array}{l} f(x, y, z, u, v) = 0 \\ \varphi(x, y, z, u, v) = 0 \end{array}\right\} \qquad (17)$$

Here we can regard three variables as independent and the other two variables as functions of these independent variables. For definiteness, let us assume that $u = u(x, y, z)$ and $v = v(x, y, z)$ and try to compute the derivatives u'_x and v'_x. For this purpose we differentiate both equations (17) (y and z are regarded as fixed). This yields

$$\left.\begin{array}{l} f'_x \cdot 1 + f'_u u'_x + f'_v v'_x = 0 \\ \varphi'_x \cdot 1 + \varphi'_u u'_x + \varphi'_v v'_x = 0 \end{array}\right\} \quad \text{i.e.} \quad \left.\begin{array}{l} f'_u u'_x + f'_v v'_x = -f'_x \\ \varphi'_u u'_x + \varphi'_v v'_x = -\varphi'_x \end{array}\right\} \qquad (18)$$

Hence, we have a system of two equations of the first degree [i.e. system (18)] containing the two unknown quantities u'_x and v'_x.

If the determinant of the system is not equal to zero the system can be solved (see Sec. VI.4). Let us therefore suppose that

$$\begin{vmatrix} f'_u & f'_v \\ \varphi'_u & \varphi'_v \end{vmatrix} \neq 0 \tag{19}$$

Solving the equations we can find the sought-for derivatives.

The derivatives with respect to y and z are found similarly. It is important that the determinant of the system for the derivatives with respect to y and z is equal to (19) again (only the right-hand sides of the system differ from the former ones). Consequently, (19) is a sufficient condition for existence of the implicit functions $u = u(x, y, z)$ and $v = v(x, y, z)$ which are defined by system of equations (17). In the general case such a condition is derived in a similar way and is analogous to (19).

A functional determinant (i.e. a determinant whose elements are the derivatives of some functions) of form (19) is widely encountered in mathematics and is called a **Jacobian** after K. Jacobi (1804-1851), a German mathematician. There is a special symbol for designating such a determinant: $\begin{vmatrix} f'_u & f'_v \\ \varphi'_u & \varphi'_v \end{vmatrix} = \dfrac{D(f, \varphi)}{D(u, v)}$. The expression $\dfrac{D(f, \varphi)}{D(u, v)}$ should be regarded as an indivisible symbol because at the present moment the denominator and the numerator taken separately make no sense to us yet.

Analogous questions arise when we have to solve a system of equations containing parameters. For example, let the following system of two equations with the two unknown quantities x and y be considered:

$$\left. \begin{array}{l} f(x, y, \alpha, \beta, \gamma, \ldots) = 0 \\ \varphi(x, y, \alpha, \beta, \gamma, \ldots) = 0 \end{array} \right\}$$

where $\alpha, \beta, \gamma, \ldots$ are parameters. Suppose that the system has the solution x_0, y_0 for certain values $\alpha_0, \beta_0, \gamma_0, \ldots$ of the parameters and $\dfrac{D(f, \varphi)}{D(x, y)} \neq 0$ for these values. Then, by the above reasoning, the system defines x and y as functions of $\alpha, \beta, \gamma, \ldots$, that is the system has a uniquely defined solution as the parameters vary and take on certain values lying in the vicinity of the values $\alpha_0, \beta_0, \gamma_0, \ldots$. By the way, as it will be shown in Sec. XII.3, the condition $\dfrac{D(f, \varphi)}{D(x, y)} \neq 0$ guarantees only the local solvability of the system because the system may not be solvable if the increments of the parameters become too large. It should be underlined that such a "stability" of a solution with respect to variations of the parameters can be guaranteed only if the number m of the equations is equal to the

number n of the unknown quantities. If $m < n$ then $n - m$ unknown quantities remain arbitrary and if $m > n$ then the solvability of the system implies that there are $m - n$ additional relationships between the parameters.

§ 4. Partial Derivatives and Differentials of Higher Orders

14. Definitions. For definiteness, let $u = f(x, y, z)$ (functions of any number of arguments can be investigated similarly). Then, as has been shown, we have the three partial derivatives of the first order $u'_x = f'_x(x, y, z)$, $u'_y = f'_y(x, y, z)$ and $u'_z = f'_z(x, y. z)$. Each of them can be differentiated repeatedly with respect to x, y and z. Hence, we obtain the following nine partial derivatives of the second order:

$$u''_{xx} = f''_{xx}(x, y, z), \ u''_{xy} = f''_{xy}(x, y, z), \quad u''_{xz} = f''_{xz}(x, y, z),$$
$$u''_{yx} = f''_{yx}(x, y, z), \ u''_{yy} = f''_{yy}(x, y, z), \quad u''_{yz} = f''_{yz}(x, y, z),$$
$$u''_{zx} = f''_{zx}(x, y, z), \ u''_{zy} = f''_{zy}(x, y, z) \text{ and } u''_{zz} = f''_{zz}(x, y, z)$$

The differentiation of elementary functions represented explicitly is performed in accordance with the rules given in Sec. IV.5. The differentiation of implicit functions is achieved by the repeated differentiation of equalities of type (14), (15) or (18) and the like. Derivatives of orders higher than the second are defined analogously.

Partial differentials of higher orders are defined in a similar manner and, just as it was done in Sec. IV.12 and Ses. 10, we arrive at the equalities

$$\partial^2_{xx} u = u''_{xx} \, dx^2, \quad \partial^2_{xy} u = u''_{xy} \, dx \, dy \quad \text{etc.} \tag{20}$$

where the differential of the independent variable x is understood as $dx = \Delta x = \partial_x x$ and so on. From this we deduce

$$u''_{xx} = \frac{\partial^2_{xx} u}{(\partial_x x)^2} = \frac{\partial^2 u}{\partial x^2}, \quad u''_{xy} = \frac{\partial^2_{xy} u}{(\partial_x x)(\partial_y y)} = \frac{\partial^2 u}{\partial x \, \partial y}, \quad u''_{xz} = \frac{\partial^2 u}{\partial x \, \partial z} \quad \text{etc.}$$

The notion of a *partial difference* of a function of several variables can be defined in a way similar to the one used in Sec. V.7. But in this case we must indicate the variable with respect to which the difference is taken. The differences can be taken with different steps for different variables. For example, let $z = f(x, y)$; then we can designate the step along the x-axis by h and use the symbol Δ_h for denoting the partial difference with respect to x: $\Delta_h z = f(x + h, y) - f(x, y)$. Similarly, let k designate the step along the y-axis and let Δ_k be the symbol for the partial difference corresponding to y: $\Delta_k z = f(x, y + k) - f(x, y)$. Then it is natural to introduce the

notation

$$\Delta_{hh}^2 z = \Delta_h (\Delta_h z), \quad \Delta_{hk}^2 z = \Delta_k (\Delta_h z) \quad \text{etc.}$$

The connection between *partial difference quotients* and the derivatives is expressed by the formulas

$$z_x' = \lim_{h \to 0} \frac{\Delta_h z}{h}, \quad z_y' = \lim_{h \to 0} \frac{\Delta_k z}{k}, \quad z_{xx}'' = \lim_{h \to 0} \frac{\Delta_{hh}^2 z}{h^2}, \quad z_{xy}'' = \lim_{\substack{h \to 0 \\ k \to 0}} \frac{\Delta_{hh}^2 z}{hk}$$

and the like.

15. Equality of Mixed Derivatives. Let $z = f(x, y)$. Then the function has four partial derivatives of the second order, namely z_{xx}'', z_{xy}'', z_{yx}'' and z_{yy}''. It turns out that the derivatives z_{xy}'' and z_{yx}'' which are called **mixed** partial derivatives are equal:

$$z_{xy}'' = z_{yx}'' \tag{21}$$

that is *the mixed derivatives are independent of the order in which the differentiation is performed*. (This is true in case z_{xy}'' and z_{yx}'' are continuous.)

To prove formula (21) it is sufficient to observe that, according to the end of Sec. 14,

$$z_{xy}'' = \lim_{h, k \to 0} \frac{1}{hk} \Delta_{hk}^2 z, \quad z_{yx}'' = \lim_{h, k \to 0} \frac{1}{hk} \Delta_{kh}^2 z \tag{22}$$

At the same time

$$\Delta_{hk}^2 z = \Delta_k (\Delta_h z) = \Delta_k [f(x + h, y) - f(x, y)] =$$
$$= [f(x + h, y + k) - f(x, y + k)] - [f(x + h, y) -$$
$$- f(x, y)] = f(x + h, y + k) - f(x, y + k) -$$
$$- f(x + h, y) + f(x, y),$$
$$\Delta_{kh}^2 z = \Delta_h (\Delta_k z) = \Delta_h [f(x, y + k) - f(x, y)] =$$
$$= [f(x + h, y + k) - f(x + h, y)] - [f(x, y + k) -$$
$$- f(x, y)] = f(x + h, y + k) - f(x + h, y) -$$
$$- f(x, y + k) + f(x, y)$$

i.e.

$$\Delta_{hk}^2 z = \Delta_{kh}^2 z$$

and hence the *mixed differences* are independent of the order in which they are calculated. From this and from (22) it follows now that formula (21) is true.

If now we consider the derivatives of orders higher than the second of a function it is permissible, in accord with formula (21), to interchange any two subsequent operations of differentiation carried out with respect to any variables. Thus, we can pass from any order of performing the differentiation to any other order. Hence, only the number of differentiations with respect to the corresponding

variables is essential here but not the order in which the operations are performed. For instance,

$$u_{xxyz}^{IV} = u_{xyxz}^{IV} = u_{xyzx}^{IV} = u_{xzyx}^{IV} = u_{zxyx}^{IV} \quad \text{etc.}$$

but at the same time these derivatives are unequal to u_{xyyz}^{IV}.

Partial differentials [see formula (20)] are also independent of the order of differentiating a function.

16. Total Differentials of Higher Order. The total differential of any given order is defined as the total differential (see Sec. 11) of the total differential of the preceding order. As before (see Sec. IV.12), the differentials of independent variables should be regarded as constant quantities in subsequent differentiations. For example. let $z = f(x, y)$. Then

$$dz = z_x' \, dx + z_y' \, dy,$$

$$d^2z = d(dz) = (z_x' \, dx + z_y' \, dy)_x' \, dx + (z_x' \, dx + z_y' \, dy)_y' \, dy =$$
$$= z_{xx}'' \, dx^2 + z_{yx}'' \, dy \, dx + z_{xy}'' \, dx \, dy + z_{yy}'' \, dy^2 =$$
$$= z_{xx}'' \, dx^2 + 2z_{xy}'' \, dx \, dy + z_{yy}'' \, dy^2 \qquad (23)$$

Here we have used formula (21). Further, we have

$$d^3z = (z_{xx}'' \, dx^2 + 2z_{xy}'' \, dx \, dy + z_{yy}'' \, dy^2)_x' \, dx +$$
$$+ (z_{xx}'' \, dx^2 + 2z_{xy}'' \, dx \, dy + z_{yy}'' \, dy^2)_y' \, dy =$$
$$= (z_{xxx}''' \, dx^3 + 2z_{xxy}''' \, dx^2 \, dy + z_{xyy}''' \, dx \, dy^2) +$$
$$+ (z_{xxy}''' \, dx^2 \, dy + 2z_{xyy}''' \, dx \, dy^2 + z_{yyy}''' \, dy^3) =$$
$$= z_{xxx}''' \, dx^3 + 3z_{xxy}''' \, dx^2 \, dy + 3z_{xyy}''' \, dx \, dy^2 + z_{yyy}''' \, dy^3$$

We see that here, as well as in deducing formula (IV.32), the calculations are performed according to the scheme in which we successively remove the brackets in the expressions $(a + b)^2$, $(a + b)^3$ etc. In the general case we have

$$d^nz = \frac{\partial^n z}{\partial x^n} \, dx^n + \binom{n}{1} \frac{\partial^n z}{\partial x^{n-1} \partial y} \, dx^{n-1} \, dy + \binom{n}{2} \frac{\partial^n z}{\partial x^{n-2} \partial y^2} \, dx^{n-2} dy^2 + \cdots$$

$$\cdots + \frac{\partial^n z}{\partial y^n} \, dy^n$$

This result can be written in the following symbolical form:

$$d^nz = \left(dx \frac{\partial}{\partial x} + dy \frac{\partial}{\partial y} \right)^n z$$

In the right-hand side of the last formula the brackets should be removed as if ∂, ∂x, ∂y, dx and dy were ordinary algebraic factors. In a similar way, if

$$u = f(x, y, z) \quad \text{then} \quad d^nu = \left(dx \frac{\partial}{\partial x} + dy \frac{\partial}{\partial y} + dz \frac{\partial}{\partial z} \right)^n u$$

and so on.

If $z = f(x, y)$ but x and y are no longer independent variables we must change formula (23) in a manner similar to the one used in the end of Sec. IV.12:

$$
\begin{aligned}
d^2z &= d\,(z'_x\,dx + z'_y\,dy) = d\,(z'_x\,dx) + d\,(z'_y\,dy) = d\,(z'_x)\,dx + \\
&\quad + z'_x d\,(dx) + d\,(z'_y)\,dy + z'_y d\,(dy) = (z''_{xx}\,dx + z''_{xy}\,dy)\,dx + \\
&\quad + z'_x\,d^2x + (z''_{yx}\,dx + z''_{yy}\,dy)\,dy + z'_y\,d^2y = \\
&= z''_{xx}\,dx^2 + 2z''_{xy}\,dx\,dy + z''_{yy}\,dy^2 + z'_x\,d^2x + z'_y\,d^2y \qquad (24)
\end{aligned}
$$

The expressions for subsequent differentials are changed in a similar way.

Solid Analytic Geometry

§ 1. Space Coordinates

1. Coordinate Systems in Space. Besides Cartesian coordinates described in Sec. VII.9, the following coordinate systems are widely used.

1. **Cylindrical coordinates** ρ, φ and z are shown in Fig. 201. To construct a cylindrical coordinate system we choose polar coordinates (in the x, y-plane) and add the third coordinate z to them.
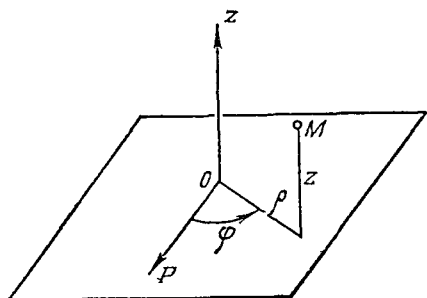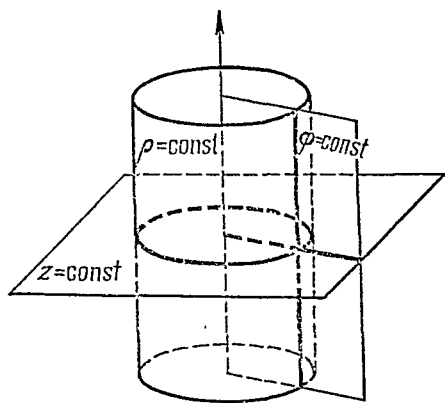
Fig. 201   Fig. 202

Obviously, for all the points in space to be described, it is sufficient that we take the following ranges for ρ, φ and z: $0 \leqslant \rho < \infty$, $-\pi < \varphi \leqslant \pi$ and $-\infty < z < \infty$.

The *coordinate surfaces*, that is surfaces on which one of the coordinates is constant whereas the other two vary, form three families of surfaces, namely $\rho = $ const, $\varphi = $ const and $z = $ const. These surfaces are depicted in Fig. 202. All these surfaces are of course regarded as being extended to infinity. The *coordinate curves*, that is curves on which two coordinates are constant whereas one of the coordinates varies, constitute three families of curves which are

formed by the intersection of the coordinate surfaces. The coordinate curves are shown in heavy lines in Fig. 202. (By the way, the coordinate surfaces of a Cartesian coordinate system are the planes parallel to the planes xOy, yOz or zOx, and the coordinate curves are the straight lines parallel to the coordinate axes Ox, Oy or Oz.)

Let us take the Cartesian coordinates (x, y, z) and the cylindrical coordinates (ρ, φ, z) which are placed as it is shown in Fig. 203.
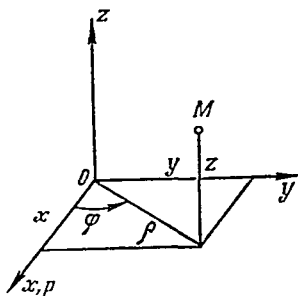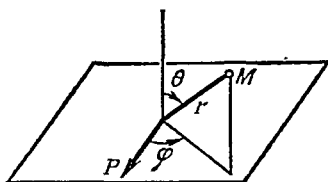
Fig. 203   Fig. 204

Then the relationship between the coordinates is expressed by the formulas $x = \rho \cos \varphi$, $y = \rho \sin \varphi$ and $z = z$.

Cylindrical coordinates are often applied to investigating solids and surfaces of revolution (such as a circular cylinder or a circular cone etc.), and the z-axis is placed along the axis of revolution in such investigations.

2. **Spherical coordinates** (which are sometimes called *spatial polar coordinates*) are shown in Fig. 204. These coordinates are analogous to the geographical coordinates. The distinction between them is that the "latitude" θ is reckoned here from "North Pole" whereas in geography it is reckoned from the equator. To describe all the points in space it is sufficient to take r, θ and φ within the limits $0 \leqslant r < \infty$, $0 \leqslant \theta \leqslant \pi$ and $-\pi < \varphi \leqslant \pi$.

The coordinate surfaces and curves are shown in Fig. 205. If Cartesian coordinates (x, y, z) and spherical coordinates (r, θ, φ) are located as it is shown in Fig. 206 then the relationship between them is expressed by the formulas

$$x = \rho \cos \varphi = r \sin \theta \cos \varphi,$$
$$y = \rho \sin \varphi = r \sin \theta \sin \varphi,$$
$$z = r \cos \theta$$

Spherical coordinates are especially convenient for investigating solids bounded by surfaces of the type shown in Fig. 205 but they are also applied in many other cases.

Cartesian, cylindrical and spherical coordinates are particular cases of the so-called **orthogonal coordinates** in which any two intersecting coordinate curves form a right angle (check up this property for the above coordinate systems!). Some non-orthogonal coordinate systems are also applied to certain problems (for instance, general affine coordinates mentioned in Sec. VII.9).

2. Degrees of Freedom. We have seen that different coordinate systems can be introduced in space. A general property common to all the systems is that the position of a point in space is specified by three coordinates whereas the position of a point in a plane is specified by two coordinates and a point on a curve by one coordinate. This property can be expressed in the following terms: there are three *degrees of freedom* when we choose a point in space (the geometric space) or when a point moves in space whereas there are two degrees of freedom when we choose a point belonging to a plane (and also to an arbitrary surface) and one degree of freedom when we consider a point on a curve. Or, in other words, the geometric space is three-dimensional whereas surfaces are two-dimensional and curves are one-dimensional.
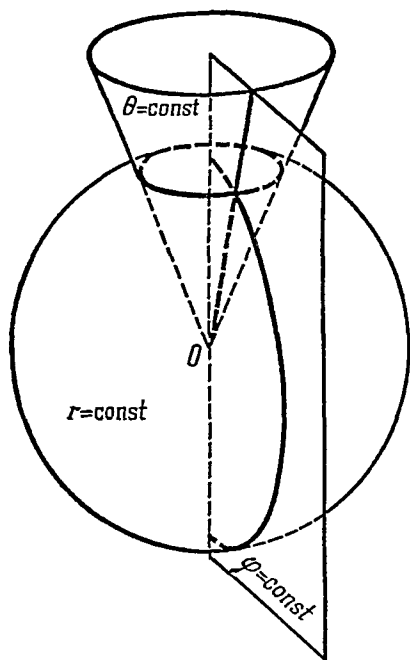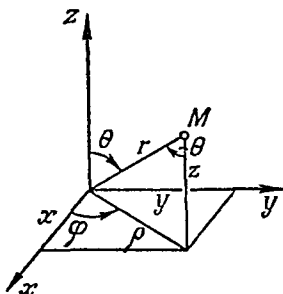
Fig. 205   Fig. 206

In the general case the notion of a degree of freedom is introduced in the following way. Let there be a certain set of objects (in the above example such a set was the totality of all the points in space). Suppose that each of the objects can be specified by indicating numerical values of some parameters (in the above example such parameters were the coordinates of a point). Let these parameters satisfy the following requirements.

(1) The parameters are independent, that is they can assume arbitrary values. For instance, if we fix all the parameters but one then this single parameter can be varied arbitrarily (sometimes within certain limits).

(2) The parameters are essential, that is any variation of the parameters leads in fact to certain changes of the object in question.

If these conditions are satisfied and if there are k such parameters then we say that we have k degrees of freedom in choosing an object from this set. The set of objects in question is then called a *k-dimensional space* (*generalized space*) or a *k-dimensional manifold*. The parameters are called coordinates (generalized coordinates) in the space. As in the case of ordinary coordinates, generalized coordinates can be chosen in different ways, and a specific choice of coordinates is usually made so that it should be convenient for the investigation. The objects which constitute a space are called its *elements* or *points*. Hence. a many-dimensional space acquires a concrete interpretation.

The above definition of a dimension is in agreement with the definition of the dimension of a linear space given in Sec. VII.19 because in such a space the coefficients of the resolution of a vector with respect to a fixed basis can be taken as parameters. But now we consider spaces which belong to a more general class. and the only connection between the elements of such a generalized space is that these objects are taken from a certain set. We can introduce the notion of closeness of the elements if we consider any elements whose parameters are close to each other as being close in the space in question. If such a notion is introduced in a space we can easily define the notion of a limit in this space. Such a space in which the notion of closeness of its elements is introduced (or. which is the same, the notion of passage to a limit is defined) is called a *topological space*.

Let us consider some examples. Let the set of all the circles lying in a plane be considered. Each of the circles is completely determined by the numerical values of three parameters, namely by the coordinates (x, y) of its centre and by its radius r. These parameters are independent (each of them can be varied arbitrarily) and essential (every variation of one of the parameters leads to a certain variation of the circle in question). Therefore, when we choose a circle in a plane we have three degrees of freedom, and hence such a set of circles is a three-dimensional generalized space with the coordinates x, y and r. Similarly, the set of all spheres in space is a four-dimensional space.

In physics we usually consider the set of events. An event is completely characterized if we can answer the questions "where does the event take place?" and "when does the event occur?". We can answer the first question by indicating the corresponding coordinates, for instance, the Cartesian coordinates x, y and z. The second question

is answered if we indicate the corresponding moment of time t. The space of events is therefore four-dimensional. and we can choose the quantities x. y. z and t as generalized coordinates in the space.

One more example: what is the number of degrees of freedom when a line segment of given length l moves in space? Each segment of this kind is completely specified by the coordinates (x_1, y_1, z_1) and (x_2, y_2, z_2) of its end-points. These coordinates can be taken as parameters which specify the position of the segment. These parameters are obviously essential but not independent since they are connected by the equation

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} = l$$

implied by formula (VII.14). Hence, only five parameters can be regarded as being independent. If we arbitrarily choose five of the parameters then the sixth parameter is expressed in terms of the five parameters with the help of the above relation. Thus, a line segment of given length has five degrees of freedom when it moves in space.

Generally, if we have n parameters which are essential but are connected by m independent equations (that is by equations such that none of them is implied by the others) then we can choose $n - m$ parameters as independent parameters, and the remaining m parameters will be expressed in terms of the former. Hence, there will be $n - m$ degrees of freedom. For instance. when a triangle moves in space we have $9 - 3 = 6$ degrees of freedom (check up the result!). This example is important in connection with the fact that the position of a rigid body ("perfectly rigid body") is completely defined if we indicate the positions of its three points which do not lie on the same straight line (why?). Consequently, we have six degrees of freedom when we investigate the motion of a rigid body in space.

In addition, let us find the number of degrees of freedom when an infinite straight line moves in space. We can reason in the following way: if we choose two arbitrary points A and B in space (each of the points is defined by its three coordinates) and draw a straight line (P) passing through the points we shall have six parameters specifying the position of the line. Since these parameters are obviously independent one may think that there are six degrees of freedom here. But such a conclusion is wrong because there are cases here when a certain variation of the parameters does not change the position of the straight line although it makes the points A and B move (along the line (P) when it occupies a fixed position). Hence, the condition that the parameters should be essential does not hold here. When the point A slides along the straight line (P) it has one

three unknowns of the form

$$
\left.
\begin{array}{l}
F_1\,(x,\ y,\ z) = 0 \\
F_2\,(x,\ y,\ z) = 0 \\
F_3\,(x.\ y,\ z) = 0
\end{array}
\right\}
$$

The notions of **algebraic** and **transcendental** sur with th
ced as it was done in Sec. II.7 for plane curves. As in rs. By
can also be singular cases here: imaginary surfaces, fact
a surface and disintegration of a surface. It should ning
account that in Sec. II.8 we considered cases when a es
degenerate into a point, but a surface can degenerate n ts
a point but also into a curve. For instance, a "sphere of zer di
is nothing but a point whereas an infinite "circular cylinder
diameter' is a straight line etc.

4. Cylinders, Cones and Surfaces of Revolution. For e
let us take the equation $z - x^2 = 0$. If we consider the cor
ding geometric figure as a plane curve then the equation rep
a parabola (L) with the equation $z = x^2$ lying in the x, z-pla
see that point O with the coordinates $x = 0$ and $z = 0$ and the
A with the coordinates $x = 2$ and $z = 4$ belong to the pa
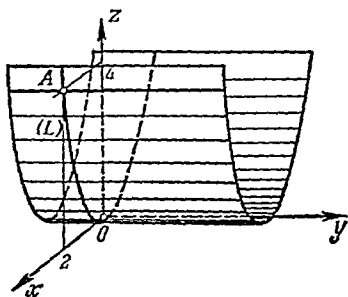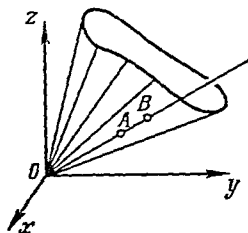
Fig. 207   Fig. 208

But we can consider the same equation with respect
spatial coordinate system x, y, z and then we obtain a cyli
surface depicted in Fig. 207 for which the relation $z - x^2$
its equation. The parabola serves as the *directing curv* of
linder, and its *elements* are parallel to the y-axis. This is a p
cylinder. Indeed, besides the point A $(2,\ 0,\ 4)$, the points $(2$
and $(2,\ -8,\ 4)$ also belong to the surface. Moreover, all the
with the coordinates $(2,\ y,\ 4)$ belong to the surface for an a
y since these coordinates satisfy the equation under consid
because the equation does not involve y and therefore y can
trary and it is only x and z that should satisfy the equation. B

We sometimes encounter the problem of finding the projection (L') of a given curve (L) lying in space on one of the coordinate planes. As an example, see Fig. 210 where the projection on the plane xOy is shown. To solve the problem we must determine the relationship between the coordinates x and y of the points belonging to (L). If (L) is represented by equations (2) then in order to find (L') we must eliminate z from the equations. If (L) is represented parametrically in form (3) then we can simply retain the first two equalities.

6. Parametric Representation of Surfaces in Space. Parametric Representation of Functions of Several Variables. It was shown in Sec. 5 that to obtain a parametric representation of a curve we must introduce one parameter [see formula (3)]. Now let us find out the geometric meaning of equations of the form

$$x = \varphi\,(u,\ v),\quad y = \psi\,(u,\ v),$$
$$z = \chi\,(u,\ v) \qquad (4)$$

containing two independent parameters u and v which can take on arbitrary numerical values. It is natural to expect that such equations represent a surface in space since a point on a surface has two degrees of freedom and therefore in order to specify such a point two parameters should be indicated.

Fig. 211

To justify this supposition let us take any two equations (4), for instance, the first two equations. Generally speaking, we can express u and v from these equations (at least theoretically) in terms of x and y. Thus we obtain expressions of the form $u = u\,(x,\ y)$ and $v = v\,(x,\ y)$. If we then substitute them into the third equation (4) we shall arrive at an equation of the form $z = z\,(x,\ y)$ which. as it was shown in Sec. IX.1, represents a surface in space. Consequently, equations (4) represent a surface in space in **parametric form**.

We shall show in Sec. XI.14 that there are some singular cases when the above transition from (4) to the formula $z = z\,(x,\ y)$ is impossible. These are the cases when the surface in question degenerates into a curve or into a point. For instance, the "surface" $x = u + v$, $y = 2u + 2v$, $z = 1 - u - v$ is in fact a line (why is it so?).

A sketch of a surface (S) represented by equations (4) is depicted in Fig. 211. If we make u assume different constant values and if we vary v then we obtain different curves lying on (S), each of these curves being completely specified by the corresponding constant

revolution in question. For example, the equation $z = ay^2$ is the equation of a parabola lying in the plane yOz. Therefore the equation $z = a(x^2 + y^2)$ is the equation of the corresponding paraboloid of revolution.

5. **Curves in Space.** A curve in space can be regarded as a line of intersection of two surfaces (that is as the locus of points common
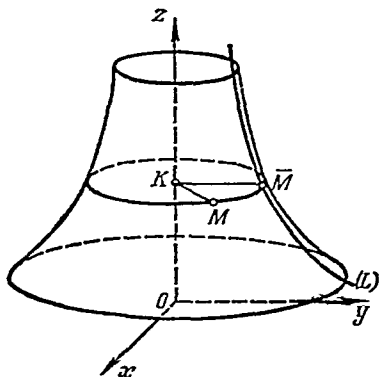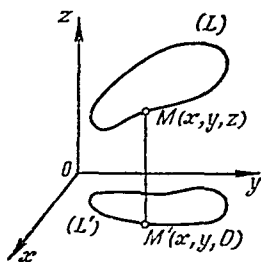
Fig. 209   Fig. 210

to both surfaces) or as a trace (trajectory) of a moving point. In the first case the equations of both surfaces can be put down in the form

$$\left. \begin{array}{l} F_1(x,\ y,\ z) = 0 \\ F_2(x,\ y,\ z) = 0 \end{array} \right\} \qquad (2)$$

Since the points of the line of intersection of the surfaces belong simultaneously to both surfaces this line is the locus of points whose coordinates simultaneously satisfy both equations (2). Hence, system (2) should be regarded as a system of two equations in three unknowns.

From the point of view of the second approach the equation of a curve has a parametric form

$$x = \varphi(t), \quad y = \psi(t), \quad z = \chi(t) \qquad (3)$$

(see Sec. VII.23). To pass from form (3) to form (2) we must eliminate t from equations (3) (for instance, we can express t in terms of x from the first equation and then substitute the result into the other two equations) if it is required and if it is possible. To perform the reverse transition from (2) to (3) (on the same conditions) we can denote one of the variables by t and then solve equations (2) for the other two variables. For instance, we can substitute $x = t$ and then solve the equations for y and z expressing these variables as functions of t.

We sometimes encounter the problem of finding the projection (L') of a given curve (L) lying in space on one of the coordinate planes. As an example, see Fig. 210 where the projection on the plane xOy is shown. To solve the problem we must determine the relationship between the coordinates x and y of the points belonging to (L). If (L) is represented by equations (2) then in order to find (L') we must eliminate z from the equations. If (L) is represented parametrically in form (3) then we can simply retain the first two equalities.

6. Parametric Representation of Surfaces in Space. Parametric Representation of Functions of Several Variables. It was shown in Sec. 5 that to obtain a parametric representation of a curve we must introduce one parameter [see formula (3)]. Now let us find out the geometric meaning of equations of the form

Fig. 211

$$x = \varphi \, (u, \, v), \quad y = \psi \, (u, \, v),$$
$$z = \chi \, (u, \, v) \qquad (4)$$

containing two independent parameters u and v which can take on arbitrary numerical values. It is natural to expect that such equations represent a surface in space since a point on a surface has two degrees of freedom and therefore in order to specify such a point two parameters should be indicated.

To justify this supposition let us take any two equations (4), for instance, the first two equations. Generally speaking, we can express u and v from these equations (at least theoretically) in terms of x and y. Thus we obtain expressions of the form $u = u \, (x, \, y)$ and $v = v \, (x, \, y)$. If we then substitute them into the third equation (4) we shall arrive at an equation of the form $z = z \, (x, \, y)$ which, as it was shown in Sec. IX.1, represents a surface in space. Consequently, equations (4) represent a surface in space in **parametric form**.

We shall show in Sec. XI.14 that there are some singular cases when the above transition from (4) to the formula $z = z \, (x, \, y)$ is impossible. These are the cases when the surface in question degenerates into a curve or into a point. For instance, the "surface" $x = = u + v$. $y = 2u + 2v$, $z = 1 - u - v$ is in fact a line (why is it so?).

A sketch of a surface (S) represented by equations (4) is depicted in Fig. 211. If we make u assume different constant values and if we vary v then we obtain different curves lying on (S), each of these curves being completely specified by the corresponding constant

value of u. This is so because when u is fixed and only v is varied we have a single parameter, i.e. only one degree of freedom. Similarly, putting $v = $ const we obtain another family of curves on (S). These curves can be taken as *coordinate curves* on (S), and the parameters u and v can be regarded as coordinates on (S).

Equations (4) define a certain functional relationship between x, y and z. Actually, if the values of x and y are given we can (at least theoretically) determine the corresponding values of u and v, as it was done above. The values u and v thus found yield the corresponding value of z. Hence, we obtain a function $z = z(x, y)$ which is originally represented in parametric form (4) and whose graph is the surface (S) considered above.

In the general case of an arbitrary number of variables the parametric representation of a functional relationship is introduced in the following way. Let there be given the equations

$$\left.\begin{array}{l} x_1 = f_1(t_1, t_2, \ldots, t_m) \\ x_2 = f_2(t_1, t_2, \ldots, t_m) \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ x_n = f_n(t_1, t_2, \ldots, t_m) \end{array}\right\} \tag{5}$$

where the variables $t_1, t_2, \ldots, t_m$ are considered to be parameters. If $m < n$ then choosing m equations we can express $t_1, t_2, \ldots, t_m$ from these equations in terms of the corresponding variables x. This is usually possible with the exception of some cases of degeneration which will be discussed in Sec. XI.14. Substituting the expressions thus obtained into the remaining equations (5) we get a representation of $n - m$ variables of type x as functions of the other m variables x. Hence, we can say that equations (5) represent an *m-dimensional manifold* lying in the n-dimensional space. When we choose a point belonging to such a manifold we have m degrees of freedom. In those cases when there is a degeneration the dimension of the manifold turns out to be less than m. If $m \geqslant n$ then, generally speaking, equations (5) do not define any functional relationship between the variables x.

The derivatives of functions represented parametrically are found by analogy with Sec. IX.13 where we differentiated implicit functions. For example, let a function $z = z(x, y)$ defined by formulas (4) be considered and let it be necessary to compute the derivative z'_x. Rewriting the first two equations (4) in the form $\varphi(u, v) - x = 0$, $\psi(u, v) - y = 0$ we can find u'_x and v'_x as it was done for equations (IX.17) (by the way, in practical calculations this can be performed without rewriting the equations). The condition guaranteeing the possibility of such computations is

$$\begin{vmatrix} \varphi'_u & \varphi'_v \\ \psi'_u & \psi'_v \end{vmatrix} \neq 0$$

Differentiating the last equation (4) we obtain the formula $z'_x = \chi'_u u'_x + \chi'_v v'_x$. Finally, substituting the values u'_x and v'_x found above into the formula we obtain the desired expression of z'_x. Starting from this result we can find the derivatives of higher order by means of repeated differentiation.

§ 3. Algebraic Surfaces of the First and of the Second Orders

7. Algebraic Surfaces of the First Order. The general form of an equation of a surface of the first order is put down as

$$Ax + By + Cz + D = 0 \tag{6}$$

(compare with Sec. II.9). To find out what surface is defined by such an equation let us introduce the vector

$$\mathbf{a} = A\mathbf{i} + B\mathbf{j} + C\mathbf{k} \tag{7}$$

Then equation (6) can be rewritten in the form $\mathbf{a} \cdot \mathbf{r} + D = 0$ [see formulas (VII.7) and (VII.12)] where $\mathbf{r}$ is the radius-vector. But $\mathbf{a} \cdot \mathbf{r} = a \, \text{proj}_a \, \mathbf{r}$ [see formula (VII.4)] which implies

$$a \, \text{proj}_a \, \mathbf{r} + D = 0, \quad \text{i.e.} \quad \text{proj}_a \, \mathbf{r} = -\frac{D}{a}$$

Thus, the surface is the locus of all points M for which the projections of their radius-vectors on the constant vector $\mathbf{a}$ have the constant value $-\frac{D}{a}$. Fig. 212 shows that this is a plane (P) which is perpendicular to the vector $\mathbf{a}$.

Hence, surfaces of the first order are planes.

Let us consider some simple problems.

(1) Investigate in what way variations of the values of the coefficients A, B, C and D affect the position of the plane (P). This can be seen from Fig. 212. For instance, if we vary D retaining some constant values of A, B and C the plane will be in a translatory motion. In particular, for $D = 0$ it passes through the origin of coordinates. Variations of the coefficients A, B and C result in rotating the vector $\mathbf{a}$ and, consequently, in rotating the plane (P). If $A = 0$ then the vector $\mathbf{a}$ lies in the plane yOz, and the plane (P) is therefore parallel to the x-axis. If, in addition, $D = 0$ the plane will pass through the x-axis.

The case when other coefficients turn into zero are investigated similarly.

Let us also put down the equations of the coordinate planes: $z = 0$ is the equation of the plane xOy, $x = 0$ is the equation of the plane yOz and $y = 0$ is the equation of the plane zOx.

in three unknowns. Each of the equations is the equation of a plane in the x, y, z-space, and hence the problem reduces to finding the point of intersection of the corresponding planes (P_1), (P_2) and (P_3). The determinant D of the system is equal to the triple scalar product of the three corresponding vectors which are perpendicular to the planes (see Sec. VII.15). If $D \neq 0$ these vectors are not parallel to the same plane, and the plane (P_3) will therefore intersect the line of intersection of the planes (P_1) and (P_2) at a single point. Hence system (VI.5) has a unique solution. If $D = 0$ the vectors are parallel to a plane T (see Fig. 214). This implies that the planes (P_1), (P_2) and (P_3) are parallel to a straight line (l), and therefore they either have no points in common at all or have infinitely many common points which constitute a straight line parallel to (l). In the first case the system has no solutions and in the second it has infinitely many solutions (the whole "straight line of solutions"). Possible dispositions of the planes are shown in Fig. 215 for both cases. (Let the reader think what other dispositions can be found here.)

8. **Ellipsoid.** We shall begin with the *canonical equation* of an ellipsoid without giving its geometric definition. The equation is of the form

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \tag{10}$$

where a, b and c are positive constants called **semi-axes** of the ellipsoid.

By analogy with Sec. II.10, we can easily verify that $| x | \leqslant a$, $| y | \leqslant b$ and $| z | \leqslant c$, that is an ellipsoid is a finite, bounded surface, that the planes xOy, yOz and zOx are its planes of symmetry and that the origin of coordinates is its centre of symmetry (the centre of the ellipsoid).

To investigate the form of an ellipsoid let us apply the *method of parallel sections.* The method consists in investigating the curves of intersection of the surface in question with the coordinate planes, that is with the planes whose equations are of the form $x = $ const, $y = $ const and $z = $ const. Let us first consider the curve of intersection of our ellipsoid with the plane $z = h$, which is parallel to the plane xOy. For this purpose let us put $z = h$ in equation (10) which results in

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 - \frac{h^2}{c^2}$$

or

$$\frac{x^2}{\left(a \sqrt{1 - \dfrac{h^2}{c^2}} \right)^2} + \frac{y^2}{\left(b \sqrt{1 - \dfrac{h^2}{c^2}} \right)^2} = 1$$

Hence, we obtain an ellipse with semi-axes $a\sqrt{1-\frac{h^2}{c^2}}$ and $b\sqrt{1-\frac{h^2}{c^2}}$. Thus, for $h=0$ we have an ellipse with semi-axes a and b. When $|h|$ is increased the ellipse decreases but remains similar to the original ellipse because the ratio of its semi-axes is constant. When h assumes the values $h=\pm c$ the ellipse degenerates into a point since its semi-axes become equal to zero. The investigation of the curves of intersection of the ellipsoid with the planes $y=h$ and $x=h$ yields similar results. Hence we conclude that the ellipsoid has the form shown in Fig. 216.

If two semi-axes are equal, for instance, if $a=b$ then the curves of intersection with the planes $z=h$ are circles. Hence, in this
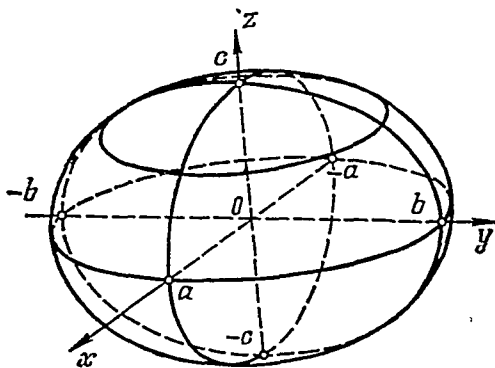
Fig. 216

case we obtain an **ellipsoid of revolution**, that is a surface generated by the revolution of an ellipse about one of its axes, instead of a **triaxial ellipsoid** which we have in the general case when the axes are unequal. An ellipsoid of revolution is also referred to as a **spheroid**. When a spheroid is generated by revolving an ellipse about its major axis it is called a **prolate ellipsoid of revolution** (it resembles an egg). If an ellipse rotates about its minor axis we have an **oblate ellipsoid of revolution**. Finally, if all the three axes are equal to each other the ellipsoid turns into a sphere.

By analogy with Sec. II.10, we can easily prove that a triaxial ellipsoid can be obtained by performing uniform contraction (or stretching) of a sphere towards two coordinate planes. The contraction towards one of the coordinate planes necessarily results in a spheroid. At the same time, Sec. XI.6 implies that when performing uniform contraction of an ellipsoid we again obtain an ellipsoid.

9. Hyperboloids. There are two types of hyperboloid. A hyperboloid **of one sheet** is represented by the canonical equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1 \qquad (11)$$

The section of the surface by the plane $z = h$ yields an ellipse whose semi-axes are $a \sqrt{1 + \frac{h^2}{c^2}}$ and $b \sqrt{1 + \frac{h^2}{c^2}}$ (check it up!). Hence, for $h = 0$ we have an ellipse with semi-axes a and b. When $|h|$ increases the sizes of the ellipse also increase and tend to infinity, as $|h| \to \infty$. All the ellipses are similar because the ratio of their semi-axes equals $\frac{a}{b} = \text{const}$. We similarly verify that the sections by the planes $y = h$ and $x = h$ are hyperbolas. Hence we obtain a surface which is depicted in Fig. 217. A hyperboloid of one sheet, like an ellipsoid, has three planes of symmetry and a centre of symmetry.

If $a = b$ we have a surface which is generated by revolving a hyperbola about its conjugate axis, that is a **hyperboloid of revolution** (of one sheet).

There is another way of interpreting a hyperboloid of one sheet. Let us first take the case when the hyperboloid of one sheet having equation (11) is a surface of revolution, that is when $a = b$. [Equation (11) turns into

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} - \frac{z^2}{c^2} = 1 \qquad (a = b)$$

for this case.] Take the plane $y = b = a$. Let us consider the section of the hyperboloid in question by this plane. For this purpose we substitute $y = b = a$ into equation (11) which results in

$$\frac{x^2}{a^2} - \frac{z^2}{c^2} = 0 \quad \text{or} \quad \left(\frac{x}{a} - \frac{z}{c}\right)\left(\frac{x}{a} + \frac{z}{c}\right) = 0$$

i.e.

$$\frac{x}{a} - \frac{z}{c} = 0 \quad \text{and} \quad \frac{x}{a} + \frac{z}{c} = 0 \quad (y = b)$$

Hence, the curve of intersection disintegrates into a pair of straight lines $\frac{x}{a} - \frac{z}{c} = 0$, $y = b$ and $\frac{x}{a} + \frac{z}{c} = 0$, $y = b$ which intersect at the point A $(0, b, 0)$ (the lines are shown in Fig. 218). Because of the axial symmetry we have the same form of section by any plane which is parallel to the z-axis and which touches the "gorge" circle (the ellipse corresponding to the section $z = h = 0$ is called the gorge ellipse; in the case $a = b$ we thus have the gorge circle). Consequently, the whole hyperboloid is entirely made up of these straight lines forming two families of straight lines, as it is shown in Fig. 218, because through each of its points there pass two straight lines which lie entirely on the surface of the hyperboloid.

This property is analogous to the properties of a cylinder or of a cone which are also made up of straight lines (their elements), but in the latter cases the straight lines belong to a single family of lines. Incidentally, we see that a hyperboloid of revolution (of one sheet) can be generated by revolving one of the two skew lines about the other. Now, passing to the general case of a hyperboloid of one sheet when the parameters a, b and c entering into (11) can take on
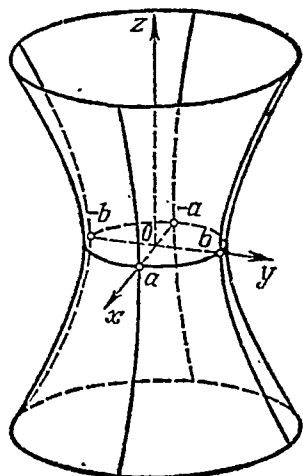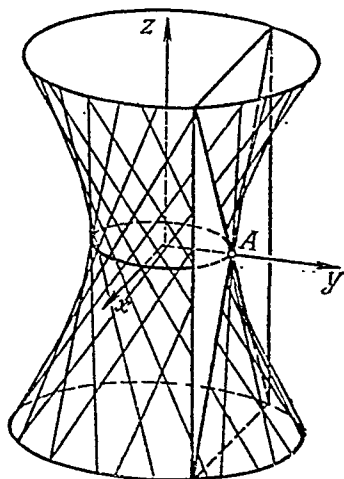
Fig. 217   Fig. 218

arbitrary positive numerical values we see that such a hyperboloid can be obtained from a hyperboloid of revolution if we uniformly contract the latter. But it is obvious that straight lines pass into straight lines under such a deformation, and therefore we conclude that a hyperboloid of one sheet of the general form is also made up of straight lines forming two families of straight lines. In conclusion, let us note that the plane depicted in Fig. 218 is the tangent plane to the hyperboloid at the point A. In fact, the tangent plane passing through a point A belonging to an arbitrary surface (S) is, by definition, a plane which touches any curve lying on (S) and passing through the point A. Therefore the tangent plane to our hyperboloid must pass through both straight lines which entirely lie on the hyperboloid and intersect at the point A. Hence, we see that a tangent plane to a surface may intersect the surface along two distinct lines.

The canonical equation of a hyperboloid of two sheets has the form

$$-\frac{x^2}{a^2} - \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

The section by the plane $z = h$ is an ellipse with semi-axes

$$a\sqrt{\frac{h^2}{c^2} - 1} \quad \text{and} \quad b\sqrt{\frac{h^2}{c^2} - 1}$$

The plane $z = h$ does not therefore intersect the surface for $|h| <$ $< c$. For $|h| = c$, that is for $h = \pm c$, we obtain ellipses whose axes are equal to zero, that is single points. When $|h|$ increases from c to infinity the sizes of the corresponding ellipses also tend to infinity. The ratio of the semi-axis being equal to the constant $\frac{a}{b}$, all the ellipses are similar. The intersection with the planes $y = h$ and $x = h$ yields hyperbolas. Hence we obtain a surface which consists of two distinct portions ("sheets") which extend to infinity. The surface is depicted in Fig. 219. If $a = b$ we obtain a hyperboloid of revolution generated by revolving a hyperbola about its transverse axis.
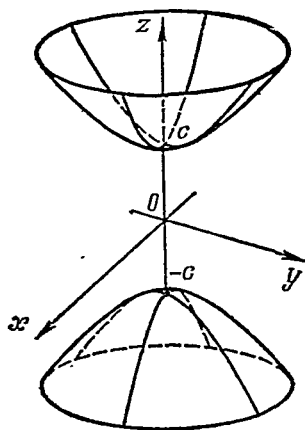
Fig. 219

10. **Paraboloids.** There are also two types of paraboloid. An **elliptic paraboloid** has the canonical equation

$$z = ax^2 + by^2 \qquad (a,\ b > 0)$$

The section by the plane $z = h$ is a curve represented by the equation $ax^2 + by^2 = h$ which can be put down as $\dfrac{x^2}{\left(\dfrac{h}{a}\right)} + \dfrac{y^2}{\left(\dfrac{h}{b}\right)} = 1$. Hence,

for $h > 0$, we obtain an ellipse with semi-axes $\sqrt{\dfrac{h}{a}}$ and $\sqrt{\dfrac{h}{b}}$.

There will be no intersection for $h < 0$, and we shall have a single point (the origin of coordinates) for $h = 0$. When h increases from 0 to infinity the sizes of the ellipses tend to infinity, and all the ellipses are similar because the ratio of their semi-axes assumes

the constant value $\sqrt{\dfrac{h}{a}} : \sqrt{\dfrac{h}{b}} = \sqrt{\dfrac{b}{a}}$. The sections by the

planes $y = h$ and $x = h$ are parabolas. Hence, we obtain a surface which is shown in Fig. 220. The paraboloid has two planes of symmetry (namely, the planes $x = 0$ and $y = 0$). If $a = b$ we have **a paraboloid of revolution** generated by revolving a parabola about its axis.

A **hyperbolic paraboloid** has the canonical equation

$$z = -ax^2 + by^2 \quad (a,\ b > 0) \tag{12}$$

Intersecting the surface by the plane $x = 0$ we obtain the parabola $z = by^2$ which opens upwards (in the positive direction of the z-axis). On the contrary, the sections by the planes $y = h$ are the parabolas $z = -ax^2 + bh^2$ which open downwards. The sections are shown in Fig. 221. Finally, the sections by the planes $z = h$ are hyperbolas. Thus, we obtain a surface having the form of a saddle.

It can be shown that this surface, like a hyperboloid of one sheet, is entirely made up of straight lines forming two families. For instance, the tangent plane to the hyperbolic paraboloid (shown
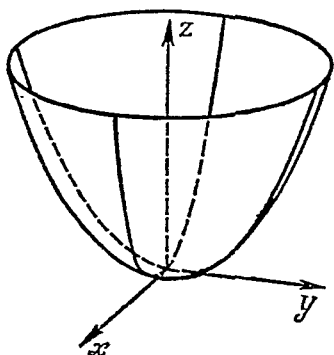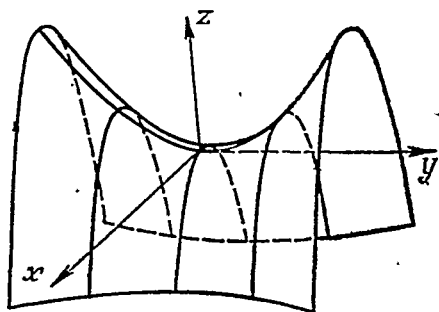
Fig. 220   Fig. 221

in Fig. 221) passing through the origin of coordinates is the plane $z = 0$. But at the same time, if we put $z = 0$ in equation (12) we get $\sqrt{a}\,x = \pm\sqrt{b}\,y$ which means that the tangent plane intersects the surface along two straight lines.

11. General Review of Algebraic Surfaces of the Second Order. The general form of an equation representing a surface of the second order can be written as

$$Ax^2 + 2Bxy + Cy^2 + 2Dxz + 2Eyz + Fz^2 + \\ + Gx + Hy + Iz + J = 0 \tag{13}$$

(compare with Sec. II.13).

We shall show in Sec. XI.12 that we can always perform a rotation of the original Cartesian axes so that the equation related to the new coordinates should no longer contain the terms involving the products of different coordinates. Therefore if we denote the new coordinates as x', y' and z' the equation will not contain the products $x'y'$, $x'z'$ and $y'z'$ and thus it will have the form

$$A'x'^2 + C'y'^2 + F'z'^2 + G'x' + H'y' + I'z' + J' = 0 \tag{14}$$

where A', C', F', G', H', I' and J' are some new constant coefficients (compare with Sec. II.13). Depending on the signs of the coefficients A', C' and F' we perform the further investigation in different ways. Let us first suppose that all these coefficients are different from zero and have the same sign. For definiteness, let them be positive. Then, as it was done in Sec. II.13, we complete the squares and perform a parallel translation of the coordinate axes which leads to a new equation of the form

$$A'x''^2 + C'y''^2 + F'z''^2 + J' = 0$$

i.e.

$$\frac{x''^2}{-\dfrac{J'}{A'}} + \frac{y''^2}{-\dfrac{J'}{C'}} + \frac{z''^2}{-\dfrac{J'}{F'}} = 1$$

where x'', y'' and z'' are the new coordinates appearing after the parallel translation. It follows that if $J' < 0$ we get the canonical equation of an ellipsoid, and therefore the original surface (13) is also an ellipsoid whose planes of symmetry and centre of symmetry are displaced and turned relative to the coordinate planes and the origin of coordinates of the original coordinate system x, y, z. But if $J' > 0$ or $J' = 0$ we obtain. respectively, an imaginary surface or a single point. Similar results are obtained for the case when the coefficients A', C' and F' are negative.

We suggest that the reader should verify that when the coefficients A', C' and F' are different from zero but have unlike signs the corresponding surface will be either a hyperboloid or a cone of the second order. We usually call such a cone elliptic although its sections, by different planes, can be ellipses, hyperbolas or parabolas (see Fig. 86). In particular, there can also be a circular cone.

If exactly one of the coefficients A', C' and F' entering into equation (14) is equal to zero, for instance, if $F' = 0$ whereas the corresponding coefficient I' is different from zero then the surface will be a paraboloid. It can be shown (but we are not going to prove it here) that in all other cases there can be only a cylinder of the second order, a pair of planes (which may coincide), degeneration into a straight line and an imaginary surface. Besides, a cylinder of the second order can have a directing curve which is an ellipse (a circle in a particular case), a hyperbola or a parabola. Accordingly, we can have an elliptic (or circular in a particular case), a hyperbolic or a parabolic cylinder. For example, the cylinder described in the beginning of Sec. 4 is parabolic.

Matrices and Their Applications

§ 1. Matrices

Matrices were first introduced by the Irish mathematician W. Hamilton (1805-1865) and the English mathematician A. Cayley (1821-1895). They are widely used now in various branches of mathematics because their application considerably simplifies the investigation of complicated systems of equations.

1. Definitions. We begin with some formal definitions whose advisability will be clarified later. A **matrix** is a rectangular array composed of numbers or some other objects. Unless the contrary is stated, we shall only deal with *real number matrices*, that is matrices composed of real numbers. For instance, such a matrix can have the form

$$\begin{pmatrix} 2 & -1.3 & 0 \\ 1 & \pi & 1 \end{pmatrix} \text{ or } \begin{pmatrix} 2 & 1 & 0 \\ -3 & \sqrt{2} & 0 \\ 2 & -1 & \frac{1}{2} \end{pmatrix} \text{ or } \begin{pmatrix} 1 \\ -2 \\ 0 \\ 3 \end{pmatrix} \text{ or } (5) \text{ etc.} \quad (1)$$

Here the parentheses are the sign of matrix. Double vertical bars are also used for this purpose (that is the notation of the form

$$\left\| \begin{matrix} 2 & 1 & 0 \\ -3 & \sqrt{2} & 0 \\ 2 & -1 & \frac{1}{2} \end{matrix} \right\|, \quad \left\| \begin{matrix} 1 \\ -2 \\ 0 \\ 3 \end{matrix} \right\|$$

etc.) but not the simple vertical bars

which designate a determinant (see § VI.1). Like in the theory of determinants, we consider elements, rows and columns of matrices. But there is an important difference between a determinant and a matrix: a determinant is equal to a certain number (see Sec. VI.1) whereas a matrix is regarded as an independent object which is not reduced to a simpler object (such as a number and the like). For brevity, we can designate a matrix by a single letter, for instance, by A, B etc., but then the letter A will nevertheless designate the

whole array of numbers. A matrix can be put down in the general
form as

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \qquad (2)$$

It is convenient to equip the elements of a matrix with two indices
under the convention that the first index indicates the number of
the row and the second the number of the column in which the
element appears. We shall sometimes use the abridged notation
$A = (a_{ij})_{mn}$ which means that i varies from 1 to m and j from 1 to n.

Every matrix is characterized by its numbers of rows and columns.
A matrix having m rows and n columns will be referred to as an
$(m \times n)$ matrix. For instance, in formulas (1) and (2) we have,
respectively, (2×3), (3×3), (4×1), (1×1), and $(m \times n)$
matrices. If the number of rows coincides with the number of co-
lumns the matrix is said to be a **square matrix**. In this case the
number of its rows and columns is called the **order** of the matrix.
A square matrix of the first order is identified with its single ele-
ment. For instance, the fourth matrix in (1) is simply the number 5.

A matrix consisting of a single column is called a **column matrix**
or a **number vector** (column-vector). Such a matrix is identified
with a vector belonging to a Cartesian space of number n-tuples
(see Sec. VII.18). Thus, the third matrix in (1) is a vector of the
space E_4, the coordinates of the vector being 1, -2, 0, 3. A matrix
having only one row is called a **row matrix**.

A matrix whose all elements are equal to zero is called a **zero
matrix**. A square matrix whose all elements are equal to zero pos-
sibly except those forming its **principal diagonal** (that is the diago-
nal connecting the left uppermost element with the right lower-
most element) is called a **diagonal matrix**. If the diagonal is formed
by the elements a, b, ..., k the diagonal matrix is denoted as
diag $(a, b, ..., k)$. If all the elements of a diagonal matrix forming
its principal diagonal are equal to unity the matrix is referred to
as a **unit matrix**. Such a matrix is usually designated by the letter **I**.
For example, the unit matrix of the third order is of the form

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \text{diag } (1, 1, 1) \qquad (3)$$

The so-called *operation of transposition* consists in interchanging
rows and columns of a matrix with the same indices. Such an ope-
ration was applied to determinants in Sec. VI.2. If we have a matrix
A then the transposed matrix (the **transpose of A**) will be designated

as A*. For instance,

$$\begin{pmatrix} 2 & \cdot 0 & 3 \\ 7 & -2 & 1 \end{pmatrix}^* = \begin{pmatrix} 2 & 7 \\ 0 & -2 \\ 3 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}^* = (1 \ 0 \ -2)$$

In the general case we can write $a_{ij}^* = a_{ji}$ (why?). Obviously, we always have $(A^*)^* = A$.

A matrix coinciding with its transpose is called **symmetric**. Of course, only a square matrix can be symmetric. The symmetry condition can be put down in the form $a_{ij} = a_{ji}$.

If we have $a_{ij} = -a_{ji}$ for all the elements of a matrix then the matrix is called **skew-symmetric (antisymmetric)**.

A square matrix A has its determinant which we shall denote by det A. For example, det $\begin{pmatrix} 1 & 0 \\ 2 & -3 \end{pmatrix} = \begin{vmatrix} 1 & 0 \\ 2 & -3 \end{vmatrix} = -3$. Only square matrices have determinants; it is impossible to speak about the determinant of a rectangular matrix which is not square. It follows from Sec. VI.2 that

$$\det I = 1 \quad \text{and} \quad \det A^* = \det A$$

2. Operations on Matrices. Two matrices with the same numbers of rows and columns are added together according to the following rule:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \end{pmatrix}$$

The multiplication of a matrix by a number is defined in an analogous way:

$$k \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} = \begin{pmatrix} ka_{11} & ka_{12} & ka_{13} \\ ka_{21} & ka_{22} & ka_{23} \end{pmatrix}$$

We can easily verify that all the axioms of linear operations (see Sec. VII.17) hold for the above operations. Hence, the set of all matrices of the same size is a linear space. Let us put down the following obvious formulas:

$$(A + B)^* = A^* + B^*, \quad (kA)^* = kA^* \quad \text{and}$$
$$\det (kC) = k^n \det C$$

where n is the order of the square matrix C. By the way, in the general case we have det $(A + B) \neq \det A + \det B$.

Now we are going to introduce the rule of multiplication of a matrix by another matrix. The advisability of this peculiar rule will be clarified in Sec. 6. First of all, for two given matrices to be

multiplied by each other, it is necessary that the number of columns of the first matrix factor be equal to the number of the rows of the second factor; if otherwise, the multiplication is impossible. This condition being fulfilled, the product is found according to the rule

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix} =$$

$$= \begin{pmatrix} a_{11}b_{11}+a_{12}b_{21} & a_{11}b_{12}+a_{12}b_{22} & a_{11}b_{13}+a_{12}b_{23} \\ a_{21}b_{11}+a_{22}b_{21} & a_{21}b_{12}+a_{22}b_{22} & a_{21}b_{13}+a_{22}b_{23} \end{pmatrix}$$

The reader should pay much attention to the structure of the formula. For instance, to obtain the element of the product belonging to the first row and to the third column we must take the first row of the first factor and the third column of the second factor and then multiply them as if we computed the scalar product of the corresponding number vector [see formula (VII.12)]. Other elements of the matrix which is the product of the two given matrices are also obtained by means of a similar operation resembling scalar multiplication of the rows of the first matrix by the columns of the second matrix. In the general case when we multiply an $(m \times n)$ matrix (a_{ij}) by an $(n \times p)$ matrix (b_{ij}) we obtain an $(m \times p)$ matrix (c_{ij}) whose elements are found according to the formula

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$$

The above rule implies that we can always mutually multiply two square matrices of the nth order which results in a square matrix of the same order. In particular, we can always multiply a square matrix by itself, that is we can raise it to the second power, but this cannot be done with a non-square rectangular matrix. There is another important particular case when we multiply a row matrix by a column matrix under the condition that they contain the same number of elements; this yields a square matrix of the first order, that is a number:

$$(a_1 \ a_2 \ a_3) \cdot \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = a_1b_1 + a_2b_2 + a_3b_3$$

By analogy with Sec. VI.2, we can verify the following properties of the product of matrices:

$$(kA)\,B = A\,(kB) = k\,(AB), \quad (A+B)\,C = AC + BC,$$
$$C\,(A+B) = CA + CB, \quad A\,(BC) = (AB)\,C$$

Of course, in all these formulas we suppose that the numbers of rows and columns of matrices entering into these expressions gua-

rantee the possibility of the corresponding multiplications. Another method of deducing the formulas will be given in Sec. 6.

The simplest examples indicate that, generally speaking, the multiplication of matrices is non-commutative, i.e. $AB \neq BA$. Let the reader verify the following relations:

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

$$(1 \quad 2) \cdot \begin{pmatrix} 0 \\ -2 \end{pmatrix} = -4, \quad \begin{pmatrix} 0 \\ -2 \end{pmatrix} \cdot (1 \quad 2) = \begin{pmatrix} 0 & 0 \\ -2 & -4 \end{pmatrix}$$

Besides, we have $\begin{pmatrix} 1 & -2 \\ -1 & -3 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ -7 \end{pmatrix}$ whereas the expression $\begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & -2 \\ -1 & -3 \end{pmatrix}$ makes no sense. The non-commutativity of matrix multiplication makes it necessary to keep the order of factors. Therefore, to specify the order, we say "to multiply A on the right by B" or simply "to multiply A by B" (the operation results in AB) but when speaking about the product BA we say "to multiply A on the left by B".

We also indicate the property

$$(AB)^* = B^*A^* \tag{4}$$

which can be easily verified, and the property

$$\det (AB) = \det A \cdot \det B \tag{5}$$

which will be proved in Sec. 7.

If A is a *complex number matrix* the symbol A^* designates the result of the operation of transposition with the simultaneous replacement of all the elements by their complex conjugates. In this case A^* is said to be the **transposed conjugate matrix**. The above formulas will hold for complex matrices if we change two of the formulas, namely if we write $\det A^* = (\det A)^*$ and $(kA)^* = k^*A^*$.

3. **Inverse Matrix.** Here we shall consider square matrices. For definiteness, let us take matrices of the third order. The role of unit matrix (3) in the operation of multiplying matrices is analogous to the role of the number 1 in the operation of multiplying numbers. Indeed, we can easily verify that $AI = IA = A$ for any matrix A.

By analogy with number multiplication, we define the notion of a matrix A^{-1} which is the inverse of the matrix A: by definition, we put

$$A^{-1}A = AA^{-1} = I \tag{6}$$

From (6) and from equality (5) it follows that

$$\det A \cdot \det (A^{-1}) = \det I = 1, \quad \text{that is} \quad \det (A^{-1}) = \frac{1}{\det A}$$

We see that for the inverse matrix to exist it is necessary that det $\mathbf{A}$ be unequal to zero: det $\mathbf{A} \neq 0$. A square matrix $\mathbf{A}$ for which det $\mathbf{A} = = 0$ is called **degenerate (singular)**. Consequently, a degenerate matrix has no inverse. At the same time, every **non-degenerate (non-singular)** matrix has its inverse. Actually, let us take an arbitrary non-degenerate matrix

$$\mathbf{K} = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} \tag{7}$$

Then, bearing in mind the definition of the product of matrices, we verify, reasoning as in Sec. VI.4, that the multiplication of $\mathbf{K}$ on the left or on the right by the matrix

$$\frac{1}{\det \mathbf{K}} \begin{pmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \\ C_1 & C_2 & C_3 \end{pmatrix} \tag{8}$$

yields the matrix $\mathbf{I}$ (the capital letters $A_1, \ldots, C_3$ designate the corresponding cofactors of the elements of the determinant of the matrix $\mathbf{K}$, see Sec. VI.3). Matrix (8) is therefore nothing but the matrix $\mathbf{K}^{-1}$.

Inverse matrices can be applied to solving matrix equations. For instance, let us consider the equation $\mathbf{AX} = \mathbf{B}$ where $\mathbf{A}$ and $\mathbf{B}$ are given matrices and $\mathbf{X}$ is an unknown matrix. Let us suppose that det $\mathbf{A} \neq 0$. Then multiplying both sides on the left by $\mathbf{A}^{-1}$ and taking advantage of equalities (6) we deduce $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}$. Similarly, the solution of the equation $\mathbf{XA} = \mathbf{B}$ is $\mathbf{X} = \mathbf{BA}^{-1}$ provided det $\mathbf{A} \neq 0$.

Matrices enable us to put down a system of equations of the first degree in an abridged form of a matrix equation. For example, system of equations (VI.5) can be rewritten in the matrix form

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}$$

(check it up!). If we designate the coefficient matrix by the letter $\mathbf{A}$, the column of unknowns (which is a number vector) by $\mathbf{x}$ and the column of the constant terms by $\mathbf{d}$ the system is put down in a still more abridged form as

$$\mathbf{Ax} = \mathbf{d} \tag{9}$$

If det $\mathbf{A} \neq 0$ then (9) implies that the solution is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{d} \tag{10}$$

If we write the formula in full we shall again obtain Cramer's rule deduced in Sec. VI.4. It should be noted that formula (9) also makes sense when the number of equations differs from the number of unknowns but in such a case the matrix A will not be square; for such a system formula (10) no longer holds because only a square matrix has its determinant.

Equation (9) with a concrete square matrix A and a column d composed of letters can be solved according to Gauss' method (see Sec. VI.5). After such a solution we come to formula (10), that is. we obtain the matrix A^{-1} as the matrix of the coefficients in the coordinates of the vector d. This method of constructing an inverse matrix is practically more convenient than the application of formula (8) especially when the order of the matrix is large.

Formula (6) indicates that the matrices A and A^{-1} are mutually inverse, that is $(A^{-1})^{-1} = A$. Besides, we sometimes apply the formula $(AB)^{-1} = B^{-1}A^{-1}$ (where $\det A \neq 0$ and $\det B \neq 0$) which can be readily verified:

$$(B^{-1}A^{-1}) (AB) = B^{-1} (A^{-1}A) B = B^{-1}IB = B^{-1}B = I$$

Finally, substituting $B = A^{-1}$ into formula (4) we deduce

$$(A^{-1})*A* = (AA^{-1})* = I* = I, \text{ that is } (A^{-1})* = (A*)^{-1}$$

4. Eigenvectors and Eigenvalues of a Matrix. Let A be a given square matrix. As we shall see later, we sometimes encounter an equation of the form

$$Ax = \lambda x \tag{11}$$

where x is an unknown number vector and λ is an unknown number, the dimension of x being equal to the order of A. Equation (11) has the *trivial solution* $x = 0$ for any λ but we shall be interested only in those λ for which the system has *non-trivial solutions*. A number λ of this kind is called an **eigenvalue** of the matrix A and the corresponding solution x of equation (11) is called an **eigenvector** of the matrix A.

Eigenvalues and eigenvectors can be found as follows. Since $x = Ix$ we can rewrite equation (11) in the form

$$(A - \lambda I) x = 0 \tag{12}$$

Comparing formula (12) with formula (9) we see that we have arrived at a system of n algebraic homogeneous linear equations in n unknowns where n is the order of the matrix A. According to Sec. VI.6, for a non-trivial solution to exist, it is necessary and sufficient that the determinant of the system be equal to zero, i.e.

$$\det (A - \lambda I) = 0 \tag{13}$$

This equation is called the **characteristic equation** of the matrix A and it enables us to find the eigenvalues λ. For instance, in the

·case of matrix (7) the equation has the form

$$\begin{vmatrix} a_1 - \lambda & b_1 & c_1 \\ a_2 & b_2 - \lambda & c_2 \\ a_3 & b_3 & c_3 - \lambda \end{vmatrix} = 0$$

Writing the determinant in full we see that this is an algebraic equation whose degree is equal to the order of the matrix A. By Sec. VIII.8, we conclude that a matrix of order n has n eigenvalues. Of course, some of them may be complex and some may coincide.

If we consider only real numbers and real vectors then equation (11) is satisfied only in the case when we take a real root of the characteristic equation (provided there are such). But if we admit complex numbers then every root of the characteristic equation can be substituted into (11).

After an eigenvalue has been found we can determine the corresponding eigenvector by solving vector equation (12). For this purpose we rewrite the equation in the form of a system of scalar equations and apply the methods of Sec. VI.6. Equation (12) implies that, for a fixed λ, the sum $y = x^1 + x^2$ of particular solutions x^1 and x^2 is a solution of the same system and that the product $z = kx$ of a solution x by a number k is also a solution. Hence, the set of all eigenvectors corresponding to a given eigenvalue is a linear subspace (see Sec. VII.18) of the space of all number vectors of dimension n.

The most important case here is when all the eigenvalues are distinct. In this case the subspace corresponding to a given eigenvalue λ is one-dimensional, that is an eigenvector corresponding to the eigenvalue is defined to within a numerical factor [here we also mean complex eigenvalues because, as it has been mentioned, characteristic equation (13) with real coefficients can have both real and imaginary roots]. The fact that such a subspace must be one-dimensional is implied by the property that nonzero eigenvectors corresponding to different eigenvalues are necessarily linearly independent and by the definition of a dimension which indicates that an n-dimensional linear space cannot have more than n linearly independent vectors. The linear independence can be proved as follows: if some eigenvectors x^1, x^2, x^3 correspond to different eigenvalues λ_1, λ_2, λ_3 and if x^1 and x^2 are linearly independent whereas $x^3 = \alpha x^1 + \beta x^2$ then multiplying both sides of the equality on the left by A we obtain $\lambda_3 x^3 = \alpha \lambda_1 x^1 + \beta \lambda_2 x^2$; after that multiplying the first equality by λ_3 and subtracting it from the second one we deduce $\alpha (\lambda_1 - \lambda_3) x^1 + \beta (\lambda_2 - \lambda_3) x^2 = 0$, which contradicts the linear independence of x^1 and x^2. The case when x^1 and x^2 are linearly dependent is treated similarly.

$(S) = (R)$ then we say that there is a mapping of the space (R) *into itself*. If every vector $\vec{y} \in (S)$ is the image of a vector $\vec{x} \in (R)$ under a mapping A then we say that A maps (R) *onto* (S).] A mapping is said to be **linear** if for any $\vec{x_1}, \vec{x_2} \in (R)$ and for any number λ we have

$$A (\vec{x_1} + \vec{x_2}) = A\vec{x_1} + A\vec{x_2} \quad \text{and} \quad A (\lambda\vec{x}) = \lambda A\vec{x} \qquad (20)$$

Applying these properties several times we readily deduce

$$A (\lambda_1\vec{x_1} + \lambda_2\vec{x_2} + \ldots + \lambda_k\vec{x_k}) = \lambda_1 A\vec{x_1} + \lambda_2 A\vec{x_2} + \ldots$$
$$\ldots + \lambda_k A\vec{x_k} \qquad (21)$$

Hence, a linear transformation does not change the form of a linear combination because the coefficients remain the same, and it is only preimages that are replaced by the corresponding images here. Hence, not only the sum of preimages goes into the sum of the images but also the difference goes into the difference and so on. Putting $\lambda = 0$ in the second equality (20) we deduce the relation $A\vec{0} = \vec{0}$ which holds for all the linear mappings. Here we have, of course, the zero vector of the space (S) on the right-hand side and the zero vector of the space (R) under the sign of the linear mapping A on the left-hand side.

For definiteness, let us suppose that the space (R) is three-dimensional and the space (S) is two-dimensional. Let us arbitrarily choose a basis $\vec{p_1}, \vec{p_2}, \vec{p_3}$ in (R) and a basis $\vec{q_1}, \vec{q_2}$ in (S). Each of the vectors $A\vec{p_j}$ belongs to (S) and it can therefore be resolved with respect to the basis $\vec{q_1}, \vec{q_2}$. Let us introduce the notation

$$\left. \begin{array}{l} A\vec{p_1} = a_{11}\vec{q_1} + a_{21}\vec{q_2} \\ A\vec{p_2} = a_{12}\vec{q_1} + a_{22}\vec{q_2} \\ A\vec{p_3} = a_{13}\vec{q_1} + a_{23}\vec{q_2} \end{array} \right\} \qquad (22)$$

Then, by formula (21), any vector $\vec{x} = x_1\vec{p_1} + x_2\vec{p_2} + x_3\vec{p_3} \in (R)$ goes into a vector $\vec{y} = A\vec{x} = y_1\vec{q_1} + y_2\vec{q_2} \in (S)$ according to the following rule:

$$\vec{y} = A (x_1\vec{p_1} + x_2\vec{p_2} + x_3\vec{p_3}) = x_1 A\vec{p_1} + x_2 A\vec{p_2} + x_3 A\vec{p_3} =$$
$$= (a_{11}x_1 + a_{12}x_2 + a_{13}x_3)\,\vec{q_1} + (a_{21}x_1 + a_{22}x_2 + a_{23}x_3)\,\vec{q_2}$$

that is

$$\left. \begin{array}{l} y_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ y_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{array} \right\} \qquad (23)$$

vectors and some other vectors by letters in ordinary type equipped with arrows.) In this case A is therefore the sign indicating the rotation of a vector; to each vector $\vec{x}$ there corresponds the vector $A\vec{x}$. In other words, we have defined a **mapping** A of the plane of vectors into itself. The terms a "transformation A" or an "operator A" are used synonymously in such a case. A given vector $\vec{x}$ is called a **pre-image** (an **original**, or an **inverse image**) and the vector $A\vec{x}$ is called the **image** of $\vec{x}$ under the mapping A.

A rotation transforming parallelograms into parallelograms, we conclude that the addition of preimages yields the addition of the corresponding images (see Fig. 222). In other words, in the case of rotation the image of a sum is the sum of the images: $A\,(\vec{x_1} + \vec{x_2}) = A\vec{x_1} + A\vec{x_2}$. We can similarly verify that the multiplication of a preimage by a number yields the multiplication of the image by the same number, i.e. $A\,(\lambda\vec{x}) = \lambda A\vec{x}$. Hence, under the mapping in question, to linear operations on the preimages there correspond analogous linear operations on images, that is linear relations between vectors remain valid after the mapping has been performed. For instance, if $\vec{x_1} = 2\vec{x_2} - 5\vec{x_3}$ then $A\vec{x_1} = 2A\vec{x_2} - 5A\vec{x_3}$ etc. This property is called the *linearity* of the mapping.

Fig. 222

$$\vec{z} = \vec{x_1} + \vec{x_2}, \quad A\vec{z} = A\vec{x_1} + A\vec{x_2},$$
$$A(\vec{x_1} + \vec{x_2}) = A\vec{x_1} + A\vec{x_2}$$

The operation of projecting all the vectors in space on a fixed plane or on a straight line possesses the same properties. The verification of the properties is left to the reader. We can take quite a different example now. Let us consider the space of all polynomials (see Sec. VII.18) and define the image of each of the polynomials as its derivative. The linearity of this mapping is implied by the facts that the derivative of a sum is equal to the sum of the derivatives and that a constant factor can be taken outside the sign of differentiation.

We now proceed to give the general definition of a linear mapping. Let two linear spaces (R) and (S) (see Sec. VII.17) be given. Suppose that there is a law, a rule, according to which to every vector $\vec{x} \in (R)$ there corresponds a certain vector $\vec{y} = A\vec{x} \in (S)$. Then we say that we are given a mapping A of the space (R) *into* the space (S). [If

$(S) = (R)$ then we say that there is a mapping of the space (R) *into itself*. If every vector $\vec{y} \in (S)$ is the image of a vector $\vec{x} \in (R)$ under a mapping A then we say that A maps (R) *onto* (S).] A mapping is said to be **linear** if for any $\vec{x}_1$, $\vec{x}_2 \in (R)$ and for any number λ we have

$$A\,(\vec{x}_1 + \vec{x}_2) = A\vec{x}_1 + A\vec{x}_2 \quad \text{and} \quad A\,(\lambda\vec{x}) = \lambda A\vec{x} \qquad (20)$$

Applying these properties several times we readily deduce

$$A\,(\lambda_1\vec{x}_1 + \lambda_2\vec{x}_2 + \ldots + \lambda_k\vec{x}_k) = \lambda_1 A\vec{x}_1 + \lambda_2 A\vec{x}_2 + \ldots$$
$$\ldots + \lambda_k A\vec{x}_k \qquad (21)$$

Hence, a linear transformation does not change the form of a linear combination because the coefficients remain the same, and it is only preimages that are replaced by the corresponding images here. Hence, not only the sum of preimages goes into the sum of the images but also the difference goes into the difference and so on. Putting $\lambda = 0$ in the second equality (20) we deduce the relation $A\vec{0} = \vec{0}$ which holds for all the linear mappings. Here we have, of course, the zero vector of the space (S) on the right-hand side and the zero vector of the space (R) under the sign of the linear mapping A on the left-hand side.

For definiteness, let us suppose that the space (R) is three-dimensional and the space (S) is two-dimensional. Let us arbitrarily choose a basis $\vec{p}_1$, $\vec{p}_2$, $\vec{p}_3$ in (R) and a basis $\vec{q}_1$, $\vec{q}_2$ in (S). Each of the vectors Ap_j belongs to (S) and it can therefore be resolved with respect to the basis $\vec{q}_1$, $\vec{q}_2$. Let us introduce the notation

$$\left.\begin{array}{l} A\vec{p}_1 = a_{11}\vec{q}_1 + a_{21}\vec{q}_2 \\ A\vec{p}_2 = a_{12}\vec{q}_1 + a_{22}\vec{q}_2 \\ A\vec{p}_3 = a_{13}\vec{q}_1 + a_{23}\vec{q}_2 \end{array}\right\} \qquad (22)$$

Then, by formula (21), any vector $\vec{x} = x_1\vec{p}_1 + x_2\vec{p}_2 + x_3\vec{p}_3 \in (R)$ goes into a vector $\vec{y} = A\vec{x} = y_1\vec{q}_1 + y_2\vec{q}_2 \in (S)$ according to the following rule:

$$\vec{y} = A\,(x_1\vec{p}_1 + x_2\vec{p}_2 + x_3\vec{p}_3) = x_1 A\vec{p}_1 + x_2 A\vec{p}_2 + x_3 A\vec{p}_3 =$$
$$= (a_{11}x_1 + a_{12}x_2 + a_{13}x_3)\,\vec{q}_1 + (a_{21}x_1 + a_{22}x_2 + a_{23}x_3)\,\vec{q}_2$$

that is

$$\left.\begin{array}{l} y_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ y_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{array}\right\} \qquad (23)$$

Thus, we have arrived at formulas which express transformation of the coordinates of a vector under a linear mapping. If we denote

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$$

then, by Sec. 2, formulas (23) can be rewritten in the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} \qquad (24)$$

The number matrix A entering into formula (24) is called the **matrix of the linear mapping (operator) A relative to the given bases** $\vec{p}_j$ and $\vec{q}_i$ since it depends not only on the mapping itself but also on the choice of the bases. The components of the number vectors x and y depend not only on the vectors $\vec{x}$ and $\vec{y}$ but also on the bases.

If we have two bases $\vec{p}_j$, $\vec{q}_i$ chosen in the spaces (R) and (S) and if it is known that any vector $\vec{x} = x_1\vec{p}_1 + x_2\vec{p}_2 + x_3\vec{p}_3 \in (R)$ is carried into a vector $\vec{y} = A\vec{x} = y_1\vec{q}_1 + y_2\vec{q}_2 \in (S)$ whose coordinates y_1, y_2 are defined by formulas (23) the mapping $\vec{y} = A\vec{x}$ is linear.

When we add vectors their similar components are also added and the corresponding number vectors are added too; but (24) implies that when we add number vectors x the corresponding vectors y are also added. The second property (20) is verified similarly.

Hence, if certain bases $\vec{p}_j$ and $\vec{q}_i$ are chosen then to every linear mapping of (R) into (S) there corresponds its matrix which is the transpose of the coefficient matrix of the expansion of the vectors $A\vec{p}_j$ with respect to the basis $\vec{q}_i$. Conversely, each matrix with the corresponding numbers of rows and columns [a (2×3) matrix in our case] is a matrix of a linear mapping of (R) into (S). Evidently, if (R) is of dimension n and (S) is of dimension m then the matrix of a mapping of (R) into (S) is an $(m \times n)$ matrix. In particular, when $(S) = (R)$ the matrix of a linear mapping is square. In such a case we usually expand vectors with respect to the same basis before and after the mapping, unless the contrary is stated.

The rank of a matrix A being equal to the maximal number of its linearly independent columns, formula (22) implies that the rank equals the number of linearly independent vectors among the vectors $A\vec{p}_j$. Thus, the rank is equal to the dimension of the linear space $A(R)$ constituted by the images of all the vectors of (R) under the mapping A. The space $A(R)$ can either coincide with (S) or constitute a subspace of (S) having a lower dimension. As it has been already mentioned, in the first case we say that (R) is mapped

onto (S). In particular, it follows that although the matrix $\mathbf{A}$ of a mapping $\mathbf{A}$ of (R) into (S) depends on the choice of bases in (R) and (S) the rank of the matrix is independent of the choice. Besides, the rank of an $(m \times n)$ matrix not exceeding n, the dimension of $\mathbf{A}(R)$ cannot exceed that of (R). Consequently, we see that a dimension cannot increase under a linear mapping (in § 4 we shall see that the same is true for a non-linear mapping).

Instead of considering a mapping of vectors into vectors we can also consider a mapping of points into points which is more visual. Let us suppose that to each point M of a plane (P) there corresponds a certain point $\overline{M}$ of a plane $(\overline{P})$. Then we can say that we are given a mapping of the plane (P) into the plane $(\overline{P})$. In such a case we shall write $\overline{M} = f(M)$ (compare with the consideration in Sec. IX.9 concerning this notation). Suppose that the mapping f is such that it does not violate rectilinearity, that is suppose that vectors lying in the plane (P) are carried into vectors lying in the plane $(\overline{P})$ under the mapping. In addition, let us suppose that equal vectors in the plane (P) go into equal vectors in the plane $(\overline{P})$. Then we can say that to each vector $\vec{x}$ of the plane (P) there corresponds a completely specified vector $\vec{y}$ of the plane $(\overline{P})$ which is independent of the disposition of the origin of $\vec{x}$ in (P). Let us denote $\vec{y}$ as $\vec{y} = \mathbf{A}\vec{x}$. Finally, let the mapping $\mathbf{A}$ be linear. (The example at the beginning of Sec. 6 satisfies all the requirements enumerated here. The planes (P) and $(\overline{P})$ coincide, and $f(M)$ is understood as the result of rotation of the point M through the angle α about the centre of rotation.)

Now let us choose an arbitrary affine coordinate system (see Sec. VII.9) with the coordinates designated as x_1, x_2, with the origin of coordinates O and the base vectors $\vec{p_1}$, $\vec{p_2}$. Then the radius-vector is represented in the form $\vec{r} = x_1\vec{p_1} + x_2\vec{p_2}$. Let us also choose in the plane $(\overline{P})$ an arbitrary affine coordinate system y_1, y_2 with the origin $\overline{O}$ and the base vectors $\vec{q_1}$, $\vec{q_2}$. Denote the coordinates of the point $f(O)$ belonging to the plane $(\overline{P})$ by b_1, b_2. Let the coordinates x_1, x_2 of a point M of the plane (P) be given. What are the coordinates y_1, y_2 of the corresponding point $f(M)$? We have

$$\overrightarrow{Of}(M) = \overrightarrow{Of}(0) + \overrightarrow{f(0)f(M)} = b_1\vec{q_1} + b_2\vec{q_2} + \mathbf{A}(\overrightarrow{OM})$$

and therefore, by above formulas for transformation of the coordinates of a vector, we obtain

$$\left.\begin{array}{l} y_1 = a_{11}x_1 + a_{12}x_2 + b_1 \\ y_2 = a_{21}x_1 + a_{22}x_2 + b_2 \end{array}\right\} \tag{25}$$

or, in the matrix notation,

$$\mathbf{y} = \mathbf{Ax} + \mathbf{b} \tag{26}$$

where $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ is the matrix of the mapping A relative to the chosen bases and $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$. Conversely, we can easily verify that if the coordinates of points are transformed according to formulas (25) the corresponding mapping possesses the properties described in the preceding paragraph. The simplest formulas are obtained when the origin of coordinates in the plane (P) is carried into the origin of coordinates in the plane $(\overline{P})$ under the mapping in question. We have $b_1 = b_2 = 0$ in such a case, and therefore formulas (25) turn into

$$\left. \begin{array}{l} y_1 = a_{11}x_1 + a_{12}x_2 \\ y_2 = a_{21}x_1 + a_{22}x_2 \end{array} \right\} \tag{27}$$

i.e. $\mathbf{y} = \mathbf{Ax}$. The coordinates of vectors are also transformed according to formulas (27) in the general case (25).

If we consider the geometric space or an abstract Cartesian space of any dimension (see Sec. VII.18) we arrive at formulas similar to (25) but with different numbers of rows and columns. The matrix form of writing will have form (26) again but in the general case the rectangular matrix A may not be square since we can have a mapping of one space into another when their dimensions are unequal.

Let us suppose now that the dimensions are equal. For simplicity's sake, let us again consider a mapping of a plane (P) into a plane $(\overline{P})$. If $\det A \neq 0$ the mapping is called affine. In this case we can multiply equality (26) on the left by A^{-1} which results in $\mathbf{x} = A^{-1}\mathbf{y} - A^{-1}\mathbf{b}$. Consequently, we obtain an equality of the same form (26). The inverse mapping of the plane $(\overline{P})$ into the plane (P) is therefore also affine.

In Fig. 223 we illustrate the most important types of affine mapping of a plane onto itself for which the origin of coordinates remains at the same place. Formulas for the transformations of coordinates and the corresponding matrices relative to a Cartesian coordinate system are also put down in Fig. 223. (Let the reader prove the formulas in the third example taking advantage of the formulas $y_1 = \rho \cos(\varphi + \alpha)$ and $y_2 = \rho \sin(\varphi + \alpha)$ where $\rho = OM = \overline{OM}$.) Of course, we can also consider different combinations of these simple mappings and additional parallel translations as well.

If $\det A = 0$ then the rank of the matrix is equal either to 1 or to 0. As it was shown above, in the first case the plane (P) is mapped onto a straight line (in particular, we have such a case when we consider the operation of projecting) and in the second case the plane goes into a point of the plane $(\overline{P})$.

By analogy with Sec. II.7, we can readily prove that the order of an algebraic curve does not change under an affine mapping. In particular, the straight lines being the only curves of the first order (see Sec. II.9), an affine mapping transforms straight lines into
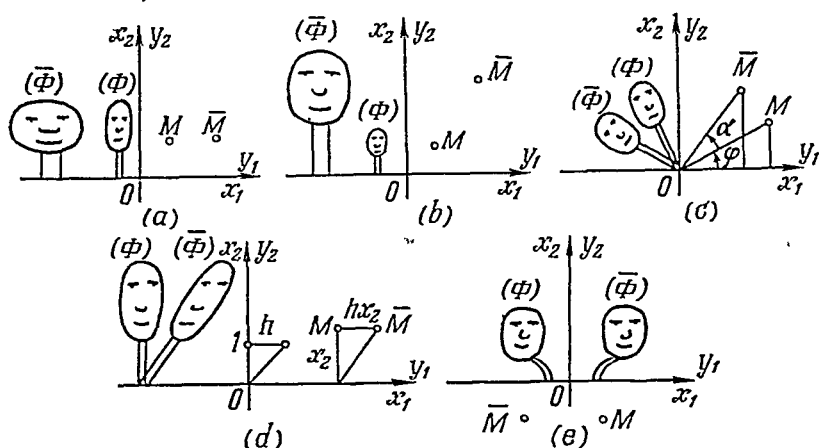
Fig. 223

Affine mappings of a plane:

(a) k-fold stretching along the x_1-axis   (b) k-fold stretching in all directions

$$y_1 = kx_1 \qquad \begin{pmatrix} k & 0 \\ 0 & 1 \end{pmatrix} \qquad\qquad y_1 = kx_1 \qquad \begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix}$$
$$y_2 = x_2 \qquad\qquad\qquad\qquad\qquad y_2 = kx_2$$

(c) Rotation through the angle α

$$y_1 = x_1 \cos\alpha - x_2 \sin\alpha \qquad \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$
$$y_2 = x_1 \sin\alpha + x_2 \cos\alpha$$

(d) Shear along the x_1-axis   (e) Reflection in the x_1-axis

$$y_1 = x_1 + hx_2 \qquad \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix} \qquad\qquad y_1 = -x_1 \qquad \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$
$$y_2 = x_2 \qquad\qquad\qquad\qquad\qquad y_2 = x_2$$

straight lines. Since the point of intersection of two lines must be carried into the point of intersection of their images under the mapping, intersecting lines go into intersecting lines, and consequently parallel straight lines go into parallel lines. On the basis of the definition we can verify that the ratio of two parallel line segments does not change under an affine mapping; at the same time, the ratio of non-parallel segments, angles and lengths are changed in the general case. Curves of the second order are carried into curves of the second order under an affine mapping. An ellipse being the only finite curve of the second order, an affine mapping transforms an ellipse into an ellipse (or into a circle in a particular case). A parabola is an infinite curve of the second order consisting of one component and it is therefore mapped onto a parabola. Finally, a hyperbola is mapped onto a hyperbola.

Let us return to linear mappings of general linear spaces (see the beginning of this section). Let two such mappings A and B of

a space (R) into a space (S) be given. Then we can define the sum of the mappings by means of the formula $(A + B)\,\vec{x} = A\vec{x} + B\vec{x}$. We similarly define the multiplication of a mapping by a number: $(\lambda A)\,\vec{x} = \lambda\,(A\vec{x})$. We readily verify that if some bases in (R) and (S) are chosen and if the above operations are performed on mappings then the same operations are performed on the matrices of the mappings. Besides, all the axioms of linear operations hold here. The role of the zero mapping is played by the mapping of the whole space (R) into the zero vector of the space (S).

The multiplication of mappings is defined as their successive performance. More exactly, suppose we have a mapping B of a space (R) into a space (S) and a mapping A of the space (S) into a space (T). Then AB is understood as a "composite" mapping of (R) into (T) which is obtained if we first perform the mapping of (R) into (S) and then perform the mapping of (S) into (T), that is $(AB)\,\vec{x} = A\,(B\vec{x})$. If some bases are chosen in (R), (S) and (T) the multiplication of the mappings yields the multiplication of their matrices which accounts for the rule of multiplication of matrices given in Sec. 2. The rule is extremely important and we shall therefore illustrate what has been said by taking an example of matrices of the second order. Let the mappings B and A be represented by the corresponding formulas

$$y_1 = b_{11}x_1 + b_{12}x_2 \atop y_2 = b_{21}x_1 + b_{22}x_2 \Big\} \quad \text{and} \quad z_1 = a_{11}y_1 + a_{12}y_2 \atop z_2 = a_{21}y_1 + a_{22}y_2 \Big\}$$

where x_j, y_k, z_i $(j, k, i = 1, 2)$ are the coordinates in the spaces under consideration. Then in order to obtain the "composite" mapping we must substitute the first formulas into the second which results in

$$z_1 = a_{11}\,(b_{11}x_1 + b_{12}x_2) + a_{12}\,(b_{21}x_1 + b_{22}x_2) =$$
$$= (a_{11}b_{11} + a_{12}b_{21})\,x_1 + (a_{11}b_{12} + a_{12}b_{22})\,x_2 = c_{11}x_1 + c_{12}x_2,$$
$$z_2 = a_{21}\,(b_{11}x_1 + b_{12}x_2) + a_{22}\,(b_{21}x_1 + b_{22}x_2) =$$
$$= (a_{21}b_{11} + a_{22}b_{21})\,x_1 + (a_{21}b_{12} + a_{22}b_{22})\,x_2 = c_{21}x_1 + c_{22}x_2$$

Thus, we have arrived at the matrix $C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$ formed according to the rule indicated in Sec. 2 which shows that $C = AB$.

In connection with the relationship between the operations on mappings and the corresponding operations on their matrices (in particular, between the operations of multiplication) we can deduce all the properties of the operations on matrices from the properties of the operations on mappings because the properties of the latter are obviously implied by their linearity. Here we shall indicate only the property $A\,(BC) = (AB)\,C$ which is implied by the corresponding property of mappings $A\,(BC) = (AB)\,C$, the latter being

justified by the fact that both on the left-hand side and on the right-hand side we have the mappings which are obtained if we first perform the mapping C and then the mappings B and A in succession. When we consider the mappings of a space into itself, that is when $(S) = (R)$, the role of the unity in the operation of multiplication is played by the identity mapping (unit mapping) under which every vector is carried into itself: $\vec{I}\vec{x} = \vec{x}$. It is clear that we always have $AI = IA = A$. The matrix of the unit mapping is the unit matrix I relative to any basis chosen in the space in question. Similarly, the inverse matrix corresponds to the inverse mapping. We saw that the multiplication of matrices is non-commutative in the general case. This is obviously accounted for by the fact that when we reverse the order in which mappings are performed the result can be changed considerably. (For instance, let the reader verify that if we first apply the mapping shown in Fig. 223a for $k = 2$ to the point $(0, 1)$ and then apply the mapping c for $\alpha = 90°$ this will result in the point $(2, 0)$. But if we reverse the order of the operations we shall obtain the point $(1, 0)$.)

Generally, two given operations, actions, are called **commuting** if the result of their successive application does not depend on the order they are performed, and they are called **non-commuting** if otherwise. (Think whether the following two "operations" commute: (a) filling a swimming-pool with water; (b) diving into the swimming-pool.)

7. Transformation of the Matrix of a Linear Mapping When the Basis Is Changed. It has been already noted that when we consider a linear mapping A of a space (R) into a space (S) the matrix of the mapping depends on the choice of the bases in both spaces. One and the same mapping can have a more complicated matrix relative to one basis and a simpler matrix relative to another basis. Let us investigate the relationship between the matrix of a mapping and the bases we choose. For this purpose let us return, for definiteness, to the example in which we deduced formula (24). Let a new basis $\vec{p_1'}, \vec{p_2'}, \vec{p_3'}$ be chosen in (R). Then any vector $\vec{x}$ can be resolved both with respect to the new basis and to the old one:

$$\vec{x} = x_1\vec{p_1} + x_2\vec{p_2} + x_3\vec{p_3} = x_1'\vec{p_1'} + x_2'\vec{p_2'} + x_3'\vec{p_3'} \tag{28}$$

where x_1, x_2, x_3 are the old coordinates and x_1', x_2', x_3' are the new coordinates of the vector $\vec{x}$. Each of the new base vectors can be expanded with respect to the old basis:

$$\left.\begin{aligned}
\vec{p_1'} &= h_{11}\vec{p_1} + h_{21}\vec{p_2} + h_{31}\vec{p_3} \\
\vec{p_2'} &= h_{12}\vec{p_1} + h_{22}\vec{p_2} + h_{32}\vec{p_3} \\
\vec{p_3'} &= h_{13}\vec{p_1} + h_{23}\vec{p_2} + h_{33}\vec{p_3}
\end{aligned}\right\} \tag{29}$$

where h_{ij} are some coefficients which define the transformation from the old basis to the new basis. Substituting formulas (29) into (28) and equalling the coefficients in the same basis vectors $\vec{p_1}, \vec{p_2}, \vec{p_3}$ we obtain

$$\left.\begin{array}{l} x_1 = h_{11}x'_1 + h_{12}x'_2 + h_{13}x'_3 \\ x_2 = h_{21}x'_1 + h_{22}x'_2 + h_{23}x'_3 \\ x_3 = h_{31}x'_1 + h_{32}x'_2 + h_{33}x'_3 \end{array}\right\} \qquad (30)$$

(let the reader verify the calculations!). It should be noted that a matrix of type

$$\mathbf{H} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix}$$

that is a *transformation matrix from new coordinates to old coordinates* (from x'_1, x'_2, x'_3 to x_1, x_2, x_3 in our concrete example) must necessarily be non-degenerate. Indeed, when the coordinates x_1, x_2, x_3 are given we must obtain certain completely specified values of x'_1, x'_2, x'_3 and system of equations (30) must therefore be compatible and must have a unique solution. Therefore, $\det \mathbf{H} \neq 0$.

Formulas (30) can be put down by analogy with formulas (23) in the abridged form

$$\mathbf{x} = \mathbf{H}\mathbf{x}' \quad \text{where} \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad \text{and} \quad \mathbf{x}' = \begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \end{pmatrix}$$

Similarly, if we introduce a new basis $\vec{q'_1}, \vec{q'_2}$ in the space (S) and if the transformation matrix from the new coordinates to the old ones is $\mathbf{K}$ then $\mathbf{y} = \mathbf{K}\mathbf{y}'$. Substituting these formulas into (24) we obtain $\mathbf{K}\mathbf{y}' = \mathbf{A}\mathbf{H}\mathbf{x}'$. Multiplying the last relation on the left by $\mathbf{K}^{-1}$ we deduce

$$\mathbf{y}' = \mathbf{K}^{-1}\mathbf{A}\mathbf{H}\mathbf{x}' \quad \text{i.e.} \quad \mathbf{y}' = \mathbf{A}'\mathbf{x}' \quad \text{where} \quad \mathbf{A}' = \mathbf{K}^{-1}\mathbf{A}\mathbf{H}$$

The matrix $\mathbf{A}' = \mathbf{K}^{-1}\mathbf{A}\mathbf{H}$ is nothing but the matrix of the mapping in question relative to the new bases.

In particular, if $(S) = (R)$ then $\mathbf{K} = \mathbf{H}$ and therefore

$$\mathbf{A}' = \mathbf{H}^{-1}\mathbf{A}\mathbf{H} \qquad (31)$$

Let us consider the geometric meaning of the determinant of the matrix of an affine mapping of a plane onto itself. Let the mapping be defined by formulas (25) and let us first suppose that the basis $\vec{p_1}, \vec{p_2}$ taken in the plane (P) is a Cartesian basis. According to our condition we have $(\bar{P}) = (P)$ here. The coordinates of vectors being

transformed according to formulas (27), the vectors $\vec{p_1}$, $\vec{p_2}$ will be carried into the vectors $\vec{s_1} = a_{11}\vec{p_1} + a_{21}\vec{p_2}$, $\vec{s_2} = a_{12}\vec{p_1} + a_{22}\vec{p_2}$, respectively (why?). The area of the square constructed on the vectors $\vec{p_1}$, $\vec{p_2}$ equals unity whereas the area of the parallelogram which is the image of the square under the mapping, that is of the parallelogram constructed on the vectors $\vec{s_1}$, $\vec{s_2}$, is equal to $\left| \begin{vmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{vmatrix} \right| =$ $= |\det A|$, according to the end of Sec. VII.13. Now remark that all the parts of a plane are changed in a similar fashion under an affine mapping and therefore the areas of all geometric figures change proportionally with the same factor of proportionality. Hence, $|\det A|$ is equal to the factor of proportionality defining the change of the areas under the mapping in question. The sign of $\det A$ also has a certain geometric meaning. Namely, if the determinant is positive the direction of describing the contour of a figure is retained under the mapping, that is if we describe the contour of a preimage in the positive direction the contour of the image is also described in the positive direction. If $\det A < 0$ then the direction of describing a contour is replaced by the opposite direction under the mapping. (Let the reader verify all the assertions for examples in Fig. 223.)

If now we pass to an arbitrary new basis we shall have

$$\det A' = \det (H^{-1}AH) = \det (H^{-1}) \det A \det H =$$
$$= (\det H)^{-1} \det A \det H = \det A$$

and thus we see that although the matrix of an affine mapping depends on the choice of the basis with respect to which it is considered the determinant of the matrix is independent of the choice, that is its geometric meaning is the same for all possible choices of the basis.

If we consider an affine mapping of one plane onto another then $|\det A|$ is also equal to the coefficient of proportionality defining the change of the areas if we measure the areas on each of the planes relative to the areas of the corresponding parallelograms constructed on the basis vectors.

In the case of an affine mapping of the geometric space onto itself we can analogously show that $|\det A|$ equals the proportionality factor defining the change of the volumes. In this case $\det A$ has the sign $+$ or $-$ depending on whether the right-handed triads of vectors remain right-handed or turn into the left-handed triads under the mapping in question. The geometric meaning of the determinant of the matrix of an affine mapping of a Cartesian space of any dimension onto itself can be interpreted in a similar way. In particular, this meaning immediately implies formula (5) because when we perform two affine mappings in succession the corresponding

factors of proportionality defining the changes of the volumes are mutually multiplied.

8. **The Matrix of a Mapping Relative to the Basis Consisting of Its Eigenvectors.** Let us consider a linear mapping A of a linear space (R) into itself. If a nonzero vector $\vec{x}$ goes into a parallel vector under the mapping, that is if the mapping of the vector $\vec{x}$ reduces to the multiplication by a scalar $(A\vec{x} = \lambda\vec{x})$ then $\vec{x}$ is called **an eigenvector of the mapping** A **corresponding to the eigenvalue** λ.

For instance, any vector parallel to the x_1-axis is an eigenvector corresponding to the eigenvalue k of the mapping shown in the first example in Fig. 223. In this example any vector parallel to the x_2-axis is also an eigenvector; it corresponds to the eigenvalue 1, that is it does not change under the mapping. Let the reader find the eigenvectors and the eigenvalues for the other examples in Fig. 223.

If we have chosen a basis in (R) we can consider the number vector x consisting of the coordinates of a vector $\vec{x}$ instead of the vector $\vec{x}$ itself. Then, according to Sec. 6, the equality defining an eigenvector acquires the form $Ax = \lambda x$, i.e. form (11). Hence, by Sec. 4, the vector x must be an eigenvector of the matrix A of the mapping in question relative to the chosen basis. In Sec. 4 we established the method of finding these vectors. On the basis of Sec. 4, we conclude that the number of the eigenvalues of the mapping A coincides with the dimension of the space (R) but there can be imaginary values and coinciding values among them. For instance, in the example (c) in Fig. 223 all the eigenvalues are imaginary (check it up!) and therefore none of the nonzero vectors remains parallel to its original direction under the mapping. (By the way, Sec. VIII.8 implies that if the dimension of (R) is odd then equation (13) possesses at least one real root and there is therefore at least one eigenvector.)

For definiteness, let us suppose that the space (R) is three-dimensional. Besides, let us suppose that there are three linearly independent (real) eigenvectors of the mapping A (let these vectors be $\vec{l}_1, \vec{l}_2, \vec{l}_3$ and the eigenvalues be $\lambda_1, \lambda_2, \lambda_3$). Let us take these vectors as a basis in (R). It turns out that in such a case the matrix A acquires a form which is especially simple. Indeed, let us write

$$\left. \begin{array}{l} y_1' = a_{11}'x_1' + a_{12}'x_2' + a_{13}'x_3' \\ y_2' = a_{21}'x_1' + a_{22}'x_2' + a_{23}'x_3' \\ y_3' = a_{31}'x_1' + a_{32}'x_2' + a_{33}'x_3' \end{array} \right\} \tag{32}$$

where the numbers a_{jk}' $(j, k = 1, 2, 3)$ are the elements of the matrix relative to the basis which are yet unknown and x_1', x_2', x_3' are the coordinates in this basis. The vector $\vec{l}_1$ has the coordinates (projec-

tions) $x_1' = 1$, $x_2' = 0$, $x_3' = 0$. After the mapping it goes into the vector $\lambda_1 \vec{l_1}$ with the coordinates $y_1' = \lambda_1$, $y_2' = 0$, $y_3' = 0$. We must therefore have

$$\lambda_1 = a_{11}'1 + a_{12}'0 + a_{13}'0,$$
$$0 = a_{21}'1 + a_{22}'0 + a_{23}'0,$$
$$0 = a_{31}'1 + a_{32}'0 + a_{33}'0$$

from which we find $a_{11}' = \lambda_1$, $a_{21}' = 0$ and $a_{31}' = 0$. We similarly obtain $a_{22}' = \lambda_2$, $a_{33}' = \lambda_3$, $a_{12}' = a_{13}' = a_{23}' = a_{32}' = 0$ (check it up!). Therefore formulas (32) in fact have the form

$$y_1' = \lambda_1 x_1', \quad y_2' = \lambda_2 x_2', \quad y_3' = \lambda_3 x_3'$$

Consequently, *the matrix of a linear mapping relative to the basis consisting of its eigenvectors has the diagonal form*

$$\mathbf{A}' = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$$

By Sec. 4, for such a basis to exist, it is sufficient that all the roots of the characteristic equation of the matrix $\mathbf{A}$ be real and distinct. Any matrix can be regarded as the matrix of a linear mapping, and the matrix of a mapping is transformed according to formula (31) when the basis is changed. Hence, the above result can also be formulated as follows: if all the roots of the characteristic equation of a square matrix $\mathbf{A}$ are real and distinct it is possible to find a non-degenerate matrix $\mathbf{H}$ such that the matrix $\mathbf{H}^{-1}\mathbf{A}\mathbf{H}$ will be diagonal and the diagonal elements will be equal to the roots.

If the characteristic equation of a matrix has an imaginary root we can find the corresponding eigenvector (number vector) by solving equations (12), and the coordinates of the vector will also be imaginary. Such an eigenvector has no geometric meaning. For instance, this is the case for the third example in Fig. 223. But of course we can use such number vectors without considering their geometric meaning. If we admit complex values of the projections in question then all the calculations remain true. In particular, the assertion of the preceding paragraph will remain true for any square matrix whose all eigenvalues (i.e. the roots of the characteristic equation) are distinct. But of course in the general case the matrix $\mathbf{H}$ may be complex.

If the characteristic equation of a matrix has multiple roots such a matrix cannot be reduced to the diagonal form in the general case. In particular, this is the case for the fourth example in Fig. 223.

In conclusion, let us note that although a matrix $\mathbf{A}$ is transformed according to formula (31) when the basis is changed its characteristic

equation does not change. Indeed, we have

$$\det(A' - \lambda I) = \det(H^{-1}AH - \lambda I) = \det[H^{-1}(A - \lambda I)H] =$$

$$= \det(H^{-1})\det(A - \lambda I)\det H = \frac{1}{\det H}\det(A - \lambda I)\det H = \det(A - \lambda I)$$

$$\tag{33}$$

which is what we set out to prove.

9. Transforming Cartesian Basis. In this section we shall suppose that (R) is not only a linear space but also a Euclidean space (see Secs. VII.20-21). Thus we can speak about Cartesian (Euclidean) bases in (R). Let us investigate the properties of a matrix which defines the transformation from a Cartesian basis to another Cartesian basis. For definiteness, let us consider the space (R) to be three-dimensional. In order to investigate the transformation we shall use formulas (29). Let $\vec{p_1}$, $\vec{p_2}$, $\vec{p_3}$ form a Cartesian basis, that is let them play the same role as vectors i, j, k in § VII.3. For $\vec{p_1'}$, $\vec{p_2'}$, $\vec{p_3'}$ also to form a Cartesian basis it is necessary and sufficient that there should be

$$\vec{p_1'}\cdot\vec{p_1'} = \vec{p_2'}\cdot\vec{p_2'} = \vec{p_3'}\cdot\vec{p_3'} = 1 \quad \text{and} \quad \vec{p_1'}\cdot\vec{p_2'} = \vec{p_2'}\cdot\vec{p_3'} = \vec{p_1'}\cdot\vec{p_3'} = 0$$

(why?). Putting down the scalar products in full according to formula (VII.12) we obtain

$$h_{11}^2 + h_{21}^2 + h_{31}^2 = h_{12}^2 + h_{22}^2 + h_{32}^2 = h_{13}^2 + h_{23}^2 + h_{33}^2 = 1,$$
$$h_{11}h_{12} + h_{21}h_{22} + h_{31}h_{32} = h_{12}h_{13} + h_{22}h_{23} + h_{32}h_{33} =$$
$$= h_{11}h_{13} + h_{21}h_{23} + h_{31}h_{33} = 0$$

These six equalities can be put down in the matrix form

$$\begin{pmatrix} h_{11} & h_{21} & h_{31} \\ h_{12} & h_{22} & h_{32} \\ h_{13} & h_{23} & h_{33} \end{pmatrix} \cdot \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

(Let the reader check up the last relation by multiplying the matrices entering into the left-hand side according to the rules of Sec. 2 and by comparing the result with the right-hand side.) Using the notation introduced in Secs. 1 and 3 we can rewrite the above condition in the form

$$H^*H = I, \quad \text{i.e.} \quad H^* = H^{-1} \tag{34}$$

A matrix satisfying these equalities, that is a matrix which is equal to its transpose, is called **orthogonal**. Hence, the property proved above can be stated as follows: the matrix of a transformation from one Cartesian coordinate system to another Cartesian system is orthogonal. We can similarly show that, conversely, if the transformation matrix is orthogonal then any Cartesian basis is necessa-

rily transformed into a Cartesian basis. [Verify that the matrices of transformations (II.3) and (II.4) are orthogonal.]

Orthogonal matrices are also encountered in connection with mappings. Namely, a linear mapping of a Euclidean space into itself is called *orthogonal* if the lengths of the vectors are not changed under the mapping (such mappings are also called isometric). Any triangle being carried into an equal triangle under such a mapping (according to the well-known test for the equality of triangles), we see that all the angles are retained in this case. An orthogonal mapping can either be a motion of the space as a whole or a combination of a motion and a reflection (in a hyperplane).

For instance, the third and the fifth mappings among those shown in Fig. 223 are orthogonal (the former defines a motion and the latter a reflection).

By analogy with formulas (22) and by arguments similar to those at the beginning of this section we can easily verify that the matrix A of an orthogonal mapping relative to a Cartesian basis is an orthogonal matrix. Conversely, if a mapping has an orthogonal matrix relative to a Cartesian basis the mapping is orthogonal.

Equality $A^*A = I$ [see formula (34)] implies that

$$\det A^* \cdot \det A = (\det A)^2 = \det I = 1$$

from which it follows that $\det A = \pm 1$. This is also implied by the geometric meaning of the determinant of the matrix of a linear mapping (see Sec. 7). The determinant equals 1 for a motion and -1 for a reflection or for a combination of a reflection and a motion.

Let us note a consequence which is used in mechanics. Let there be given a motion of the geometric space for which the origin of a coordinate system remains at the same place. Such a motion can be regarded as an orthogonal mapping A of the totality of all the vectors of a three-dimensional space. If we factor the left-hand side of the characteristic equation according to formula (VIII.25) in the form

$$\det (A - \lambda I) = -(\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3)$$

and then put $\lambda = 0$ in this identity we receive $\lambda_1\lambda_2\lambda_3 = 1$. It follows that at least one of the eigenvalues λ_k is real and positive. Hence, there exists a vector $\vec{x_0} \neq \vec{0}$ for which $A\vec{x_0} = \lambda_k\vec{x_0}$. But the lengths being retained, we must have $\lambda_k = 1$. Consequently, the motion in question is a rotation about an axis passing through the origin of coordinates and parallel to an eigenvector corresponding to an eigenvalue 1.

10. Symmetric Matrices. Here we shall indicate the application of the above results to investigating symmetric matrices (see Sec. 1).

It is possible to prove the following properties of the symmetric matrices.

All the eigenvalues of a symmetric matrix are real.

For example, a symmetric matrix of the second order is of the form

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} \tag{35}$$

and its characteristic equation is

$$\begin{vmatrix} a-\lambda & b \\ b & c-\lambda \end{vmatrix} = 0$$

or

$$\lambda^2 - (a+c)\,\lambda + ac - b^2 = 0 \tag{36}$$

The roots of the equation are

$$\lambda_{1,2} = \frac{a+c}{2} \pm \sqrt{\frac{(a+c)^2}{4} - ac + b^2} = \frac{a+c}{2} \pm \sqrt{\frac{(a-c)^2}{4} + b^2} \tag{37}$$

and they are obviously real. Here we shall not give the proof of this assertion and of the two following assertions in the general case for matrices of order higher than the second.

Eigenvectors of a symmetric matrix corresponding to different eigenvalues are necessarily orthogonal to each other.

As an example, let us take matrix (35) for $b \neq 0$. The coordinates of an eigenvector are found from system (12) which has the form

$$\begin{aligned} (a-\lambda)\,x_1 + bx_2 &= 0 \\ bx_1 + (c-\lambda)\,x_2 &= 0 \end{aligned}\Bigg\}$$

in our case. If λ is an eigenvalue these two equations are dependent (see Sec. VI.6) because the determinant

$$\begin{vmatrix} a-\lambda & b \\ b & c-\lambda \end{vmatrix}$$

equals zero. Hence, we can limit ourselves to solving only one of the equations. For definiteness, let us take the first equation. In order to satisfy the equation we can put $x_1 = -b$ and $x_2 = a - \lambda$. Putting $\lambda = \lambda_1$ and $\lambda = \lambda_2$ we thus obtain two eigenvectors of the form

$$\begin{pmatrix} -b \\ a-\lambda_1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -b \\ a-\lambda_2 \end{pmatrix} \tag{38}$$

The scalar product of the vectors is equal to

$$\begin{aligned} b^2 + (a-\lambda_1)\,(a-\lambda_2) &= \lambda_1\lambda_2 - a\,(\lambda_1 + \lambda_2) + b^2 + a^2 = \\ &= ac - b^2 - a\,(a+c) + b^2 + a^2 = 0 \end{aligned}$$

[in deducing the result we have taken the advantage of the well-known formulas for the sum and for the product of the roots of

quadratic equation (36)]. This implies the perpendicularity of vectors (38). If we put $b = 0$ the vectors $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ can serve as eigenvectors (verify this!) and thus we see that in this case they are also mutually perpendicular.

If not all the roots of the characteristic equation are distinct, that is if there is at least one multiple root (let the multiplicity of the root be k), it is possible to find k mutually orthogonal eigenvectors corresponding to this eigenvalue.

For example, formula (37) implies that matrix (35) possesses a double eigenvalue if and only if $a = c$ and $b = 0$. But then all the vectors are eigenvectors (check it up!) and thus we can choose two mutually perpendicular vectors among them.

The above properties imply that if a given symmetric matrix **A** is regarded as the matrix of a linear mapping relative to a Cartesian basis then we can always find a new Cartesian basis which entirely consists of eigenvectors of the matrix **A**. For instance, in the three-dimensional case the characteristic equation is of degree three and thus it has three roots which, as it has been indicated, will be real. If these roots are distinct from each other the corresponding eigenvectors are mutually perpendicular. We can choose these vectors so that they should be of unit length, and then they can be taken as the sought-for basis. If $\lambda_1 = \lambda_2 \neq \lambda_3$ we can choose two mutually perpendicular eigenvectors corresponding to the eigenvalue λ_1, and the vector corresponding to the eigenvalue λ_3 will be perpendicular to both vectors. Finally, if all the three eigenvalues are the same we can indicate three mutually perpendicular eigenvectors corresponding to the eigenvalue.

The transformation from one Cartesian basis to another is performed by means of an orthogonal matrix (see Sec. 9). Besides, if the latter basis consists of eigenvectors of a matrix it takes the diagonal form after being transformed according to formula (31) (see Sec. 8). Consequently, the property proved in the preceding paragraph can be formulated in terms of matrices as follows: for any symmetric matrix **A**, it is possible to find an orthogonal matrix **H** such that the matrix **H⁻¹AH** is diagonal and the diagonal elements are equal to the eigenvalues of the matrix **A**.

§ 3. Quadratic Forms

11. Quadratic Forms. A quadratic form in several variables is a homogeneous polynomial of the second degree in these variables. For example. a quadratic form in the three variables x_1, x_2, x_3 has the general form

$$F = a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + 2a_{23}x_2x_3 \quad (39)$$

23*

FACULTY OF ENGIN...

where $a_{11}, a_{22}, \ldots, a_{23}$ are numerical coefficients [some of the coefficients are doubled in (39) to simplify further formulas]. The symmetric matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}$$

is called the **matrix of the quadratic form**. With the help of the matrix we can rewrite formula (39) as

$$F = (a_{11}x_1 + a_{12}x_2 + a_{13}x_3)\, x_1 + (a_{12}x_1 + a_{22}x_2 + a_{23}x_3)\, x_2 +$$
$$+ (a_{13}x_1 + a_{23}x_2 + a_{33}x_3)\, x_3 = y_1x_1 + y_2x_2 + y_3x_3 =$$
$$= (x_1 \ x_2 \ x_3) \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

where

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{12}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{13}x_1 + a_{23}x_2 + a_{33}x_3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Thus, if we introduce the number vector $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ we obtain

$$F = x^*Ax \qquad (40)$$

Conversely, if a form is represented as (40) and if the matrix A is symmetric then A is the matrix of this quadratic form.

Let us now perform an arbitrary linear transformation of the variables of form (30). The transformation can be put down in the matrix form as

$$x = Hx' \qquad (41)$$

Then, by formula (4), we have $x^* = x'^*H^*$ which implies

$$F = x'^*H^*AHx' = x'^* (H^*AH)\, x'$$

that is

$$F = x'^*A'x' \quad \text{where} \quad A' = H^*AH \qquad (42)$$

But the matrix A' is symmetric because, by formula (4), we have

$$A'^* = (H^*AH)^* = H^*A^*H^{**} = H^*AH = A'$$

Hence, it is A' that is the matrix of the quadratic form after the change of variables is made.

Thus, substitution (41) yields the transformation of the matrix of a quadratic form according to formula (42). In particular, if H is an orthogonal matrix then, by formula (34), we see that $A' = H^{-1}AH$. As it has been shown (see Sec. 10), we can always choose

a matrix H such that there should be $A' = \text{diag}\,(\lambda_1, \lambda_2, \lambda_3)$ where the diagonal elements are the eigenvalues of the matrix A. But then the quadratic form will acquire the diagonal form

$$F = \lambda_1 x_1'^2 + \lambda_2 x_2'^2 + \lambda_3 x_3'^2 \tag{43}$$

in the new variables. Consequently, any quadratic form (39) can be reduced to diagonal form (43) (where λ_1, λ_2, λ_3 are the eigenvalues of the matrix A) by means of transformation (30) with an orthogonal matrix H.

The above formal transformation has the following geometric meaning. Let us regard A as the matrix of a linear mapping A relative to a Cartesian basis with the coordinates x_1, x_2, x_3. Then transformation (41) reducing the quadratic form F to form (43) corresponds to the transformation to a new basis consisting of eigenvectors of the mapping A.

In Sec. 8 [see formula (33)] we showed that any transformation of form (31) does not change the determinant $\det\,(A - \lambda I)$. Hence, if we expand the determinant in powers of λ the coefficients in the powers will not change; they will be invariant with respect to any transformation of a Cartesian coordinate system to another one. For instance, a quadratic form in two variables is expressed by the formula

$$Ax^2 + 2Bxy + Cy^2$$

(we have put down the formula using the notation applied in analytic geometry), that is its matrix is of the form

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix}$$

and its characteristic equation is written as

$$\begin{vmatrix} A-\lambda & B \\ B & C-\lambda \end{vmatrix} = \lambda^2 - (A+C)\,\lambda + AC - B^2 = 0$$

Hence, the expressions $A + C$ and $AC - B^2$ are invariant with respect to any change of Cartesian coordinates (see Sec. II.13).

12. Simplification of Equations of Second-Order Curves and Surfaces. The transformation of a quadratic form described in Sec. 11 is applied, in particular, to simplifying equations of curves and surfaces of the second order. Let us dwell on equations of surfaces since the problem of simplifying equations of curves of the second order was considered in Sec. II.13.

Let the equation of a second-order surface be represented in ordinary form (X.13) used in analytic geometry. The transformation to a new Cartesian coordinate system having the same origin is

reduced to a change of variables of the form

$$\left.\begin{array}{l} x = h_{11}x' + h_{12}y' + h_{13}z' \\ y = h_{21}x' + h_{22}y' + h_{23}z' \\ z = h_{31}x' + h_{32}y' + h_{33}z' \end{array}\right\} \qquad (44)$$

as it was shown in Sec. 9, where $\mathbf{H} = (h_{ij})$ is an orthogonal transformation matrix. (It is evident that if the origin of coordinates is left unchanged the coordinates of points and the coordinates of vectors are transformed according to the same formulas.) Substituting these expressions into equation (X.13) we see that the groups of summands containing the terms of the first degree and of the second degree are transformed independently. Let us consider the transformation of the group of the second-order terms which is a quadratic form. On the basis of Sec. 9, we conclude that we can always choose a coordinate system x', y', z' so that this group of terms should acquire the diagonal form

$$\lambda_1 x'^2 + \lambda_2 y'^2 + \lambda_3 z'^2$$

Hence, the whole equation (X.13) will have the form

$$\lambda_1 x'^2 + \lambda_2 y'^2 + \lambda_3 z'^2 + G'x' + H'y' + I'z' + J = 0 \qquad (45)$$

where λ_1, λ_2, λ_3 are the roots of the equation

$$\begin{vmatrix} A-\lambda & B & D \\ B & C-\lambda & E \\ D & E & F-\lambda \end{vmatrix} = 0$$

and G', H', I' are some new coefficients in the terms of the first degree which occur after substitution (44) has been made. Equation (45) is nothing but equation (X.14) put down in the different notation. It was investigated in Sec. X.11.

§ 4. Non-Linear Mappings

13. General Notions. Let us begin with a mapping of a plane into a plane. Suppose that there are two planes (P) and $(\overline{P})$ (by the way, the planes may coincide). Let to each point M of the plane (P) (or to each point M taken from a domain in the plane) there correspond a point $\overline{M}$ of the plane $(\overline{P})$, according to a certain law. Then we say that we are given a mapping of the plane (P) (or of its domain) into the plane $(\overline{P})$. For the mappings of a specific class, curves go into curves and geometric figures into geometric figures under a mapping of the plane (P) into the plane $(\overline{P})$ although the form of a geometric figure may change considerably (see Sec. VIII.11). Such a mapping is depicted in Fig. 224, and we see that

wing a preimage in the plane (P) it is difficult to recognize its
ge in the plane $(\overline{P})$ and vice versa. There are also the cases of
neration when some geometric figures are "contracted" into curves
even into points.
; is sometimes necessary to consider the inverse mapping which
be obtained if we arbitrarily choose images in $(\overline{P})$ and find the
esponding preimages in (P). As in Sec. I.21, where we conside-
inverse functions, it can happen that we encounter a difficulty,
iely the fact that the inverse of a single-valued mapping may
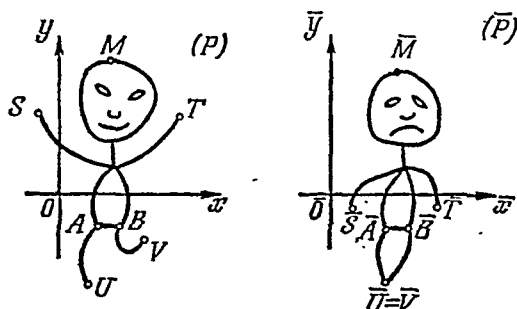
Fig. 224

be single-valued. This will be the case if two distinct points in
plane (P) (for instance, such points as the points U and V in
. 224) are carried into the same point of the plane $(\overline{P})$. Such a
at will have at least two preimages.
f not only the mapping in question but also its inverse mapping
single-valued we say that there is a **one-to-one mapping**. If
mapping under consideration is not one-to-one but does
degenerate the plane (P) can be broken into parts such
t the mapping is one-to-one in each of the parts.
lappings can be described analytically by means of coordinate
tems. To do this suppose that there is a Cartesian coordinate
tem x, y in the plane (P) and a system $\overline{x}$, $\overline{y}$ in the plane $(\overline{P})$.
ase systems may also be coincident. Then if we set the coordinates
y of a point M the coordinates $\overline{x}$, $\overline{y}$ of the corresponding point $\overline{M}$
l be completely specified. In other words, the mapping is defined
some relationships of the form

$$\overline{x} = \overline{x}(x, y), \quad \overline{y} = \overline{y}(x, y) \tag{46}$$

ien considering the inverse mapping we set the values of $\overline{x}$ and
n these formulas and find x and y. For the mapping to be one-
one. it is necessary that there should be no more than one solution
y of equations (46) for any given x and y.

Similarly, a mapping of a three-dimensional space into another three-dimensional space is defined by equations of the form

$$\overline{x} = \overline{x}\,(x,\ y,\ z), \quad \overline{y} = \overline{y}\,(x,\ y,\ z), \quad \overline{z} = \overline{z}\,(x,\ y,\ z) \qquad (47)$$

in place of (46).

We can also consider mappings for spaces of different dimensions. For instance, the formulas

$$\overline{x} = \overline{x}\,(x,\ y), \quad \overline{y} = \overline{y}\,(x,\ y), \quad \overline{z} = \overline{z}\,(x,\ y)$$

define a mapping of a plane into a three-dimensional space.

Formulas (X.5) can also be regarded as formulas defining a mapping of an m-dimensional space with the coordinates $t_1, t_2, \ldots, t_m$ into an n-dimensional space with the coordinates $x_1, x_2, \ldots, x_n$. Of course, the coordinate systems may not be Cartesian in the general case.

14. Non-Linear Mapping in the Small. Let us consider a mapping defined by formulas (46) in the vicinity of a point $M_0\,(x_0,\ y_0)$ which is mapped into a point $\overline{M}_0\,(\overline{x}_0,\ \overline{y}_0)$. The increment of any function being close to its differential to within the terms of higher order of smallness (see Sec. IX.11), we can neglect these terms and put down

$$\left.\begin{array}{l} \Delta\overline{x} = \left(\dfrac{\partial\overline{x}}{\partial x}\right)_0 \Delta x + \left(\dfrac{\partial\overline{x}}{\partial y}\right)_0 \Delta y \\[3mm] \Delta\overline{y} = \left(\dfrac{\partial\overline{y}}{\partial x}\right)_0 \Delta x + \left(\dfrac{\partial\overline{y}}{\partial y}\right)_0 \Delta y \end{array}\right\} \qquad (48)$$

Here $\Delta\overline{x} = \overline{x} - \overline{x}_0$, $\Delta x = x - x_0$ etc. (we can say that these are Cartesian coordinates reckoned from $\overline{M}_0$ and M_0, respectively) and the index "zero" indicates that the derivatives are taken at the point M_0. Comparing these formulas with formulas (27) we conclude that a non-linear mapping can be regarded as a linear mapping in an infinitesimal neighbourhood of any point with an accuracy of infinitesimals of higher order of smallness.

On the basis of Sec. 6, we conclude that if the determinant

$$\begin{vmatrix} \dfrac{\partial\overline{x}}{\partial x} & \dfrac{\partial\overline{x}}{\partial y} \\[3mm] \dfrac{\partial\overline{y}}{\partial x} & \dfrac{\partial\overline{y}}{\partial y} \end{vmatrix} = \dfrac{D\,(\overline{x},\ \overline{y})}{D\,(x,\ y)} \qquad (49)$$

(see the notation in Sec. IX.13) is unequal to zero at a point M_0 the mapping in question is one-to-one in an infinitesimal neighbourhood of the point. Moreover, it can even be regarded as affine with an accuracy of infinitesimals of higher order. Besides, the absolute value of the determinant is equal to the proportionality factor

defining the change of the areas of infinitesimal geometric figures (placed infinitely close to the point) under the mapping. The coefficient is no longer constant in the whole plane as it was·in the case of a linear mapping because the determinant takes on different values at different points in the general case. In particular, the meaning of the determinant makes it possible to attribute a certain geometric meaning separately to the denominator and to the numerator of the expression $\dfrac{D(\bar{x}, \bar{y})}{D(x, y)}$. Namely, we can consider them as being equal to the areas of an infinitesimal figure before the mapping and after the mapping, respectively.

If Jacobian (49) vanishes at a point then the mapping in question degenerates at the point, namely, the area of an infinitesimal geometric figure becomes an infinitesimal of higher order of smallness after the mapping has been performed. Finally, if Jacobian (49) is identically equal to zero the mapping degenerates throughout the whole plane which leads to a reduction of the dimension: the plane can be mapped into a line (not necessarily into a straight line) or even into a point.

One must not think that in case Jacobian (49) does not turn into zero at all the points of a finite domain the mapping in question will be one-to-one in the domain. A mapping can be non-degenerate at all the points and nevertheless it may not be one-to-one (see Fig. 225).

A mapping of a three-dimensional space into another three-dimensional space possesses similar properties. Such a mapping can be defined by formulas (47). Here the value of the Jacobian $\dfrac{D(\bar{x}, \bar{y}, \bar{z})}{D(x, y, z)}$ is also essential. Its absolute value is equal to the factor of proportionality defining the changes of the volumes of infinitely small solids. (What is the geometric meaning of the sign of the Jacobian?)

If the Jacobian is identically equal to zero we can pose the problem of determining the "degree" of the degeneration, that is whether the space x, y, z will be mapped onto a surface or onto a curve (or even into a point) of the space $\bar{x}, \bar{y}, \bar{z}$. The answer to the question is implied by the considerations given in Sec. 6 and by the fact that every mapping is linear in the small (to within infinitesimals of higher order). Thus, we must investigate the matrix

$$\begin{pmatrix} \dfrac{\partial \bar{x}}{\partial x} & \dfrac{\partial \bar{x}}{\partial y} & \dfrac{\partial \bar{x}}{\partial z} \\[2mm] \dfrac{\partial \bar{y}}{\partial x} & \dfrac{\partial \bar{y}}{\partial y} & \dfrac{\partial \bar{y}}{\partial z} \\[2mm] \dfrac{\partial \bar{z}}{\partial x} & \dfrac{\partial \bar{z}}{\partial y} & \dfrac{\partial \bar{z}}{\partial z} \end{pmatrix} . \tag{50}$$

In the case of degeneration its rank is less than three at any point (why?). If it is equal to 2 everywhere (except for some points at which the rank can be reduced still lower) then the x, y, z-space is mapped onto a two-dimensional surface. If the rank does not exceed unity at any point but does not vanish identically then the space is mapped onto a one-dimensional curve. Finally, if it is identically
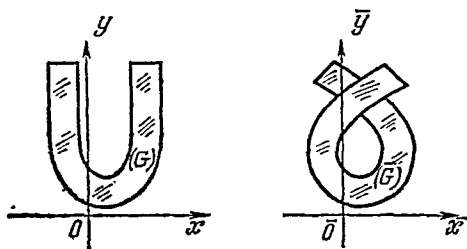
Fig. 225

equal to zero [this means that all the elements of matrix (50) are identically equal to zero] the whole space will be mapped into a point because this can be only if $\bar{x}$, $\bar{y}$, $\bar{z}$ are identically constant.

A similar situation occurs when we consider mappings of spaces of arbitrary dimensions which may be different in the general case. As it has already been said, formulas (X.5) can be regarded as formulas defining a mapping of an m-dimensional space with the coordinates t_1, t_2, ... t_m into an n-dimensional space with the coordinates x_1, x_2, ..., x_n. To determine the dimension of the manifold which appears as a result of the mapping we must compose an $(n \times m)$ matrix of the form

$$\begin{pmatrix} \dfrac{\partial x_1}{\partial t_1} & \dfrac{\partial x_1}{\partial t_2} & \cdots & \dfrac{\partial x_1}{\partial t_m} \\[2ex] \dfrac{\partial x_2}{\partial t_1} & \dfrac{\partial x_2}{\partial t_2} & \cdots & \dfrac{\partial x_2}{\partial t_m} \\[1ex] \cdot & \cdot & \cdots & \cdot \\[1ex] \dfrac{\partial x_n}{\partial t_1} & \dfrac{\partial x_n}{\partial t_2} & \cdots & \dfrac{\partial x_n}{\partial t_m} \end{pmatrix}$$

If the rank of the matrix equals k (we admit a further reduction of the rank at some points) then the dimension is equal to k.

15. Functional Relation Between Functions. The results of Sec. 14 can be applied to the notion of "functionally dependent" systems of functions. Let us first suppose that we are given three functions of three independent variables:

$$F_1(x, y, z), \qquad F_2(x, y, z), \qquad F_3(x, y, z) \tag{51}$$

We say that the functions are *dependent* on each other if there is a relation of the form

$$\Phi\,(F_1\,(x,\ y,\ z),\quad F_2\,(x,\ y,\ z),\quad F_3\,(x,\ y,\ z)) \equiv 0 \qquad (52)$$

connecting the functions where Φ is a function of three variables of the form $\Phi = \Phi\,(\lambda,\ \mu,\ \nu)$ which is not identically equal to zero. If otherwise, we say that the functions are *independent*. We can solve relation (52) for one of the variables F_1, F_2 or F_3, and therefore we can also say that functions (51) are dependent if one of them is expressible as a function of the others.

For example, the functions

$$\left(\frac{x-y}{x+z}\right)^2, \quad \ln\,(x+z) \quad \text{and} \quad x-y \qquad (53)$$

are dependent because if we denote them as F_1, F_2, F_3 we have $F_1 e^{2F_2} - F_3^2 \equiv 0$.

To establish a test for the existence of a functional relation between functions in the general case let us consider an auxiliary mapping of the form

$$\left.\begin{aligned} \lambda &= F_1\,(x,\ y,\ z) \\ \mu &= F_2\,(x,\ y,\ z) \\ \nu &= F_3\,(x,\ y,\ z) \end{aligned}\right\} \qquad (54)$$

For functions (51) to be dependent, it is necessary that relation (52), that is the relation $\Phi\,(\lambda,\ \mu,\ \nu) = 0$, should be true for all x, y, z. The relation defines a surface in the λ, μ, ν-space. Therefore, we see that for functions (51) to be dependent, it is necessary that the x, y, z-space be carried into a surface in the λ, μ, ν-space under mapping (54). This means that the mapping should be degenerate.

Applying the result of Sec. 14 we arrive at the condition

$$\frac{D\,(F_1,\ F_2,\ F_3)}{D\,(x,\ y,\ z)} \equiv 0 \qquad (55)$$

which is necessary and sufficient for functions (51) to be dependent. [Let the reader check up the fulfilment of condition (55) for functions (53).]

If the rank of the matrix

$$\begin{pmatrix} \dfrac{\partial F_1}{\partial x} & \dfrac{\partial F_1}{\partial y} & \dfrac{\partial F_1}{\partial z} \\[2mm] \dfrac{\partial F_2}{\partial x} & \dfrac{\partial F_2}{\partial y} & \dfrac{\partial F_2}{\partial z} \\[2mm] \dfrac{\partial F_3}{\partial x} & \dfrac{\partial F_3}{\partial y} & \dfrac{\partial F_3}{\partial z} \end{pmatrix}$$

equals unity the x, y, z-space, as it was shown in Sec. 14, is carried into a curve in the λ, μ, ν-space. Taking advantage of equations (X.2) we then conclude that in this case the functions F_1, F_2, F_3

are connected with each other by two independent relations of form (52).

A similar result is also obtained in the case when the number of independent variables differs from the number of functions. For instance, two functions of three variables of the form $F_1(x, y, z)$ and $F_2(x, y, z)$ will be dependent if the mapping

$$\lambda = F_1(x, y, z), \quad \mu = F_2(x, y, z)$$

will transform the x, y, z-space into a curve with an equation of the form $\Phi(\lambda, \mu) = 0$ in the λ, μ-plane. According to Sec. 14, the condition guaranteeing the above property is that the rank of the matrix

$$\begin{pmatrix} \dfrac{\partial F_1}{\partial x} & \dfrac{\partial F_1}{\partial y} & \dfrac{\partial F_1}{\partial z} \\[2mm] \dfrac{\partial F_2}{\partial x} & \dfrac{\partial F_2}{\partial y} & \dfrac{\partial F_2}{\partial z} \end{pmatrix}$$

should be less than two, that is there should be equalities of the form

$$\begin{vmatrix} \dfrac{\partial F_1}{\partial x} & \dfrac{\partial F_1}{\partial y} \\[2mm] \dfrac{\partial F_2}{\partial x} & \dfrac{\partial F_2}{\partial y} \end{vmatrix} \equiv 0, \quad \begin{vmatrix} \dfrac{\partial F_1}{\partial x} & \dfrac{\partial F_1}{\partial z} \\[2mm] \dfrac{\partial F_2}{\partial x} & \dfrac{\partial F_2}{\partial z} \end{vmatrix} \equiv 0 \quad \text{and} \quad \begin{vmatrix} \dfrac{\partial F_1}{\partial y} & \dfrac{\partial F_1}{\partial z} \\[2mm] \dfrac{\partial F_2}{\partial y} & \dfrac{\partial F_2}{\partial z} \end{vmatrix} \equiv 0$$

The condition for an arbitrary number of functions of any number of arguments to be functionally related is put down in a similar form. It should be noted that in case the number of functions exceeds the number of independent variables the functions are always dependent.

Applications of Partial Derivatives

§ 1. Scalar Field

1. Directional Derivative. Gradient. Let a Cartesian coordinate system x, y, z in space be given. Then, according to Sec. IX.9, a stationary scalar field can be regarded as a function $u = u(x, y, z)$. (When investigating a non-stationary field we can apply the same point of view at any fixed moment of time.) Besides, let a point M in space also be given. Suppose that a curve (L) starts from the point M in the direction l (see Fig. 226). Then the rate of change of the field in this direction (related to unit length) is called the **derivative of u along the direction l:**

$$\frac{\partial u}{\partial l} = \lim_{\Delta s \to 0} \frac{u(N) - u(M)}{\Delta s} \quad (1)$$
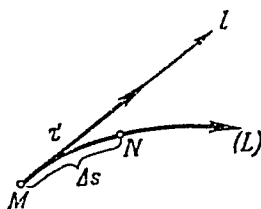
Fig. 226

To compute the directional derivative let us suppose that the curve (L) is represented in parametric form by the equation $\mathbf{r} = \mathbf{r}(s)$ where the parameter s is the arc length reckoned along (L) (see Sec. VII.23). Then the values of u taken along (L) form a composite function of the arc length: $u(s) = u(x(s), y(s), z(s))$. The sought-for derivative is then nothing but the derivative $\frac{du}{ds}$. Therefore, by the rule of differentiating a composite function (IX.11), we have

$$\frac{\partial u}{\partial l} = \frac{\partial u}{\partial x}\frac{dx}{ds} + \frac{\partial u}{\partial y}\frac{dy}{ds} + \frac{\partial u}{\partial z}\frac{dz}{ds}$$

The right-hand side can be represented as a scalar product of two vectors [see formula (VII.12)]:

$$\frac{\partial u}{\partial l} = \left(\frac{\partial u}{\partial x}\mathbf{i} + \frac{\partial u}{\partial y}\mathbf{j} + \frac{\partial u}{\partial z}\mathbf{k}\right) \cdot \left(\frac{dx}{ds}\mathbf{i} + \frac{dy}{ds}\mathbf{j} + \frac{dz}{ds}\mathbf{k}\right)$$

The first of the vectors is called the **gradient** of the field (function) u. It is designated as

$$\operatorname{grad} u = \frac{\partial u}{\partial x} \mathbf{i} + \frac{\partial u}{\partial y} \mathbf{j} + \frac{\partial u}{\partial z} \mathbf{k} \tag{2}$$

The meaning of this vector will be discussed a little later. The second vector

$$\frac{dx}{ds} \mathbf{i} + \frac{dy}{ds} \mathbf{j} + \frac{dz}{ds} \mathbf{k} = \frac{d(x\mathbf{i} + y\mathbf{j} + z\mathbf{k})}{ds} = \frac{d\mathbf{r}}{\partial s} = \boldsymbol{\tau}$$

is the unit vector in the direction l (see Sec. VII.23). Thus,

$$\frac{\partial u}{\partial l} = \operatorname{grad} u \cdot \boldsymbol{\tau} \tag{3}$$

The first factor entering into the right-hand side depends only on the choice of the point M. The second factor depends only on the choice of the direction l. In particular, we see that $\frac{du}{dl}$ is independent of the choice of a concrete curve (L) among all the possible curves passing through M in the given direction l. (By the way, it should be noted that the derivative $\frac{\partial^2 u}{\partial l^2}$ will no longer be independent of the choice.)

According to formula (VII.5), we deduce from (3) the expression

$$\frac{\partial u}{\partial l} = \operatorname{proj}_l (\operatorname{grad} u) = \operatorname{grad}_l u \tag{4}$$

($\operatorname{grad}_l u$ designates the projection of the gradient on the axis passing in the direction l).

Note that the derivatives u_x', u_y' and u_z' are also directional derivatives: for instance, u_x' is the derivative in the direction of the x-axis.

Let us put down one more useful formula containing the gradient which is based on definition (IX.7) of the total differential:

$$du = \frac{\partial u}{\partial x} dx + \frac{\partial u}{\partial y} dy + \frac{\partial u}{\partial z} dz =$$

$$= \left(\frac{\partial u}{\partial x} \mathbf{i} + \frac{\partial u}{\partial y} \mathbf{j} + \frac{\partial u}{\partial z} \mathbf{k} \right) \cdot (dx\mathbf{i} + dy\mathbf{j} + dz\mathbf{k}) =$$

$$= \operatorname{grad} u \cdot d(x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) = \operatorname{grad} u \cdot d\mathbf{r}$$

Let a field u and a point M be given. Let us set the following problem: in what direction l is the derivative $\frac{\partial u}{\partial l}$ maximal? We see that on the basis of formula (4) the problem reduces to the following question: in what direction is the projection of the vector $\operatorname{grad} u$ maximal? Evidently, the maximal projection of any vector is obtained when we take its own direction, the maximal projection being equal to the modulus of the vector.

Thus, the vector grad u at a point M indicates the direction of the maximal rate of increase of the field (function) u, this maximal rate (related to unit length) being equal to $|\operatorname{grad} u|$. The faster the change of the field, the greater the modulus. See Fig. 227 where the outer circle bounds a part of a heat conducting medium which is being cooled from outside and which is being heated from the internal region (shaded). The arrows represent the vector field of gradients of the scalar temperature field in question. We see that the gradient of the temperature is directed "toward the stove".

The physical meaning of the gradient implies that the relationship between a scalar field and its gradient is invariant, that is it remains the same when an original Cartesian coordinate system is replaced by another because the rate and the direction of maximal increase of a field are independent of the choice of a coordinate system. [By the way, the original definition (2) of the gradient which is connected with a particular choice of a Cartesian coordinate system does not directly imply the invariance.] Moreover, if we are given a field u we can find the direction and the rate of maximal increase of the field u at every point in space and hence we can find the vector grad u without using coordinates and without representing the field as a function $u(x, y, z)$. Thus, vectors grad u form a completely specified *vector field of gradients* corresponding to a given scalar field.
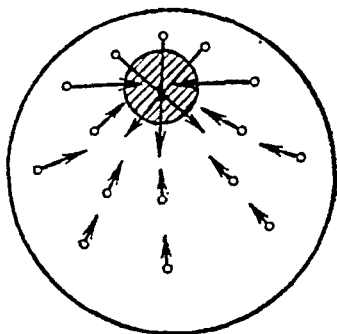
Fig. 227

Analogous conditions of invariance are set for all basic notions of the theory of vector field which we will not study in this chapter. The matter is that when we change an original Cartesian coordinate system the projections of vectors change although the vectors themselves remain invariant. Therefore, if a concept related to the theory of vector field is formulated in terms of coordinates or projections of a field we must additionally verify whether the concept satisfies the condition of invariance with respect to the changes of coordinates and projections when the coordinate axes are rotated.

Let us illustrate the application of the concept of gradient to the problem of computing the rate of change of a scalar field along a trajectory. Suppose we have a field u which may be non-stationary in the general case, that is $u = u(x, y, z, t)$. Besides, let a certain law of motion of a particle M be given in the form $\mathbf{r} = \mathbf{r}(t)$. If we consider the value of u at M in the process of motion then the value becomes a composite function of time: $u = u(x(t), y(t), z(t), t)$. To compute the sought-for rate of change of the value we

can apply transformations similar to the ones given above. This leads to the so-called **total derivative**

$$\frac{du}{dt} = \text{grad } u \cdot \mathbf{v} + \frac{\partial u}{\partial t} = \frac{\partial u}{\partial \tau} v + \frac{\partial u}{\partial t}$$

where $\mathbf{v} = \frac{d\mathbf{r}}{dt}$ is the velocity vector of the particle in the process of its motion and $\frac{\partial u}{\partial \tau}$ is the derivative of u in the direction of the tangent to the trajectory.

In case the field is stationary, that is if $\frac{\partial u}{\partial t} \equiv 0$, we have only the first summand on the right-hand side. Hence, this summand represents the rate of change of the field which is due only to the transition of the point M from one value of u to another along the trajectory. For instance, if u is temperature such a summand describes the changes of the temperature which are due to the transition of the point M from one region in space to another region with different temperature and the like. This is the so-called **convective velocity**. The second summand represents the rate of change at a motionless point (coinciding with the current position of the moving point M at a certain moment of time) which is due to the non-stationarity of the field. This is the local velocity. In the general case we have both factors which add together and yield the resultant rate of change of the field along the trajectory which is the sum of the convective velocity and the local velocity.

2. **Level Surfaces.** Level surfaces of a field $u(x, y, z)$ (see Sec. IX.7) are the surfaces on which the field assumes constant values, that is the surfaces represented by equations of the form $u(x, y, z) = \text{const.}$ Depending on the physical meaning of the field in question these surfaces may be called isothermic surfaces, isobaric surfaces and the like. There is a simple relationship between these surfaces and the gradient of the field: at each point M the gradient is normal (i.e. perpendicular to the tangent plane) to the level surface passing through the point M.

Actually, as it is seen from Fig. 228, the surfaces $u = C$ and $u = C + \Delta C$ can be regarded as being almost plane near the point M if ΔC is sufficiently small, and besides $\frac{\partial u}{\partial l} \approx \frac{\Delta u}{\Delta s} = \frac{\Delta C}{\Delta s}$. But it is clear that if l is directed along the normal to the surface the quantity Δs will assume its least value, and $\frac{\partial u}{\partial l}$ will therefore assume its maximal value. This implies our assertion.

In particular, we see that the assertion enables us to solve the following problem: to find the equation of the tangent plane passing through a point $M_0(x_0, y_0, z_0)$ of a surface (L) having an equation of the form $F(x, y, z) = 0$. To solve the problem let us introduce

a scalar field in space by means of the equation $u = F(x, y, z)$. Then (L) becomes one of the level surfaces of the field because we have $u = F(x, y, z) = 0$ on the surface. Then the vector

$$(\operatorname{grad} u)_{M_0} = \left(\frac{\partial F}{\partial x}\right)_0 \mathbf{i} + \left(\frac{\partial F}{\partial y}\right)_0 \mathbf{j} + \left(\frac{\partial F}{\partial z}\right)_0 \mathbf{k}$$

(the subscript "zero" indicates that the corresponding derivatives are taken at the point M_0) is perpendicular to the sought-for tangent plane. Hence, according to Sec. X.7 (see problem 2), we obtain the equation of the plane:

$$\left(\frac{\partial F}{\partial x}\right)_0 (x - x_0) + \left(\frac{\partial F}{\partial y}\right)_0 (y - y_0) + \left(\frac{\partial F}{\partial z}\right)_0 (z - z_0) = 0 \qquad (5)$$

The last equation can be put down as $dF = 0$. Let the reader think how we could deduce this equation in a direct way.

A surface for which the tangent plane is to be constructed can be represented by an equation of the form $z = f(x, y)$. Here we can
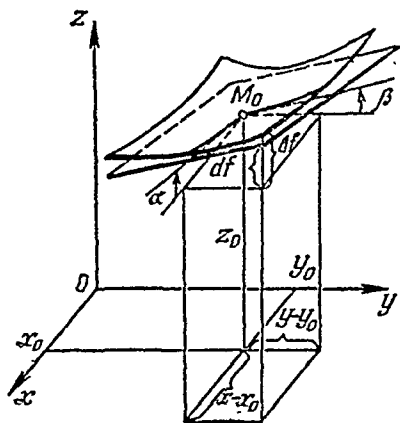
Fig. 228

Fig. 229

$$\tan \alpha = \left(\frac{\partial f}{\partial x}\right)_0, \qquad \tan \beta = \left(\frac{\partial f}{\partial y}\right)_0$$

rewrite the equation as $z - f(x, y) = 0$ and denote its left-hand side by $F(x, y, z)$. Then formula (5) is directly applicable, and thus we have $-\left(\frac{\partial f}{\partial x}\right)_0 (x - x_0) - \left(\frac{\partial f}{\partial y}\right)_0 (y - y_0) + (z - z_0) = 0$, i.e.

$$z - z_0 = \left(\frac{\partial f}{\partial x}\right)_0 (x - x_0) + \left(\frac{\partial f}{\partial y}\right)_0 (y - y_0) \qquad (6)$$

The right-hand side being equal to the total differential df, we thus obtain the geometric meaning of the total differential of a function of two independent variables. Namely, the differential is equal to the increment of the third coordinate of the point in the tangent plane (see Fig. 229).

Take an example. Let us compute the gradient of a **centrally symmetric** field $u = f(r)$ where $r = |\mathbf{r}| = \sqrt{x^2 + y^2 + z^2}$. In

this case the level surfaces are concentric spheres with centre at the origin of coordinates (why is it so?). If we take two spheres for which the difference of their radii is equal to dr then the difference of the corresponding values of the function f which are taken on these surfaces will be equal to df. Therefore, the rate of change of the function in a direction which is transversal to the level surfaces (that is along a radius) is equal to $\frac{df}{dr}$. Hence,

$$\text{grad } u\,(r) = \frac{df}{dr}\,\mathbf{r}^0 = \frac{1}{r}\,\frac{df}{dr}\,\mathbf{r} \tag{7}$$

where $\mathbf{r}^0 = \frac{\mathbf{r}}{r}$ is the unit vector in the direction of the vector $\mathbf{r}$. [Let the reader obtain result (7) on the basis of definition (2).]

3. Implicit Functions of Two Independent Variables. Implicit functions of two arguments were discussed in Sec. IX.13. Now we can approach them from a new point of view. Let us consider the equation

$$F\,(x,\,y,\,z) = 0 \tag{8}$$

in the vicinity of the point $M_0\,(x_0,\,y_0,\,z_0)$ at which the equation is satisfied. The equation defines a surface (L) in space passing through the point M_0. If $\left(\frac{\partial F}{\partial z}\right)_0 \neq 0$ [see condition (IX.16)] then, by formula (2), the vector $(\text{grad } F)_{M_0}$ has a nonzero component in the direction of the z-axis. This implies that the tangent plane to (L) passing through M_0 [which is perpendicular to $(\text{grad } F)_{M_0}$] is not parallel to the z-axis. Therefore, near the point M_0 at which the surface (L) touches the plane the angle between the z-axis and the surface (L) is different from the right angle. Hence, in the vicinity of M_0 equation (8) defines a relationship of the form $z = z\,(x,\,y)$. This functional relationship is local (i.e. it is defined only near M_0 or, as we say, "in the small") because if we take a point which is not sufficiently close to M_0 we can encounter the case when there are several values of z corresponding to given values of x and y or when there are no such values at all (see Fig. 230). It should be noted that the condition for the existence of a system of implicit functions established in Sec. IX.13 is also of a local character because it guarantees their existence only in the vicinity of the point in question.

If $\left(\frac{\partial F}{\partial z}\right)_0 = 0$ then the tangent plane to (L) is parallel to the z-axis at the point under consideration (as it is at the point N_0 in Fig. 230). In such a case it can happen that even for some points $(x,\,y)$ lying very close to N_0 equation (8) does not define a one-valued function $z = z\,(x,\,y)$. For instance, we see that some values of x and y taken near the point N_0 yield two possible values of z but at the same time there are no such values at all for other values of

x and y because the surface is convex in the direction of the radius at the point N_0. But if, for example, we have $\dfrac{\partial F}{\partial y} \neq 0$ at N_0 then near N_0 equation (8) defines a function $y = y(x, z)$. Equation (8) can happen not to define any coordinate as a function of the other coordinates near a point belonging to a surface (L) represented by an equation $F(x, y, z) = 0$ only if we simultaneously have

$$\frac{\partial F}{\partial x} = 0, \quad \frac{\partial F}{\partial y} = 0, \quad \frac{\partial F}{\partial z} = 0 \tag{9}$$

at this point.

Points at which conditions (9) hold are called **singular points** of the surface (L). A "typical" point belonging to a surface defined by
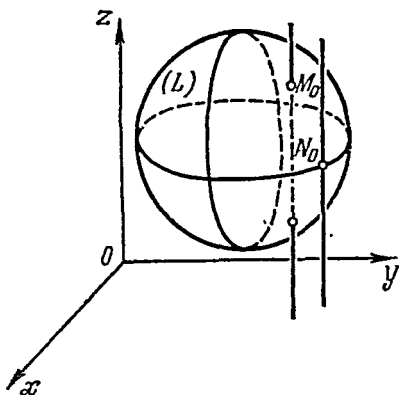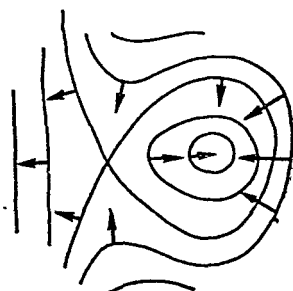
Fig. 230   Fig. 231

equation (8) is not singular because such points are found by solving the system of four equations (8) and (9) in three unknowns x, y and z which is inconsistent (overdetermined) in the general case. Therefore most of the surfaces have no singular points. Among the well-known surfaces only conic surfaces possess singular points which are their vertices.

4. Plane Fields. All the notions established for space fields are transferred with corresponding simplifications to plane fields (see the end of Sec. IX.9). For instance, the gradient $\operatorname{grad} u = \dfrac{\partial u}{\partial x}\mathbf{i} + + \dfrac{\partial u}{\partial y}\mathbf{j}$ of the field $u(x, y)$ is a vector lying in the x, y-plane. The gradient of a plane field is normal to the level line, that is to the curve represented by an equation of the form $u(x, y) = \mathrm{const}$, at each point (x, y) (see Fig. 231). In this case the meaning of the gradient implies that its modulus is approximately inversely proportional to the distance between

the level lines, that is the level lines are closer to each other in those regions where the gradient is longer.

Naturally, the equation of the tangent line to a curve represented by an equation of the form

$$f(x, y) = 0 \tag{10}$$

is obtained from (5) by dropping the third summand.

Equation (10) locally defines a function $y = y(x)$ if $f'_y \neq 0$. Singular points of curve (10) are the points for which

$$f'_x = 0 \quad \text{and} \quad f'_y = 0 \tag{11}$$

Let us introduce a surface (L) having the equation $z = f(x, y)$. Then curve (10) can be interpreted as the line of intersection of (L) by the plane $z = 0$. If conditions (10) and (11) are fulfilled at a point
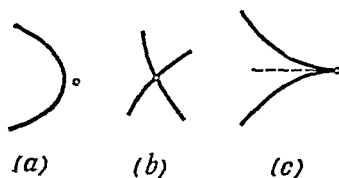
Fig. 232
(a) Isolated singular point
(b) Nodal point (c) Cusp
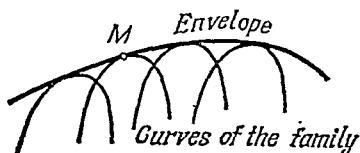
Fig. 233

formula (6) implies that the plane is tangent to the surface (L) at the point. In Sec. 9 we shall investigate the form of the line of intersection of a surface with its tangent plane near the point of tangency. We shall see that usually singular points of a plane curve are isolated points, nodal points (double points) or, more seldom, cusps (see Fig. 232).

5. Envelope of One-Parameter Family of Curves. Let us consider a family of curves dependent on a single parameter C (a *one-parameter family*). The general form of an equation of such curves can be put down as

$$F(x, y, C) = 0 \tag{12}$$

Making C assume a certain concrete value we isolate an individual curve from the family. It often happens that the disposition of the curves resembles Fig. 233. In such a case we say that the family possesses an **envelope**, that is a curve (which usually does not enter into the family) which touches some curve of the family at each of its points. To find the equation of an envelope we note that each of its points belongs to a curve of the family and equation (12) is therefore satisfied at each point. But, at the same time, in

moving along the envelope the value of the quantity C [determining the curve of the family which the envelope touches at a point (x, y)] varies: $C = C (x)$. Differentiating equality (12) with respect to x (after the ordinate of the envelope which is a function of x has been substituted for y) we obtain

$$F'_x + F'_y \; y'_{envelope} + F'_C \; C'_x = 0 \qquad (13)$$

where $y'_{envelope}$ is the slope of the envelope (i.e. of its tangent line) at an arbitrary point M. But, the envelope touching a curve of the family at the point M, the slope of the envelope equals the slope of the curve, that is $y'_{curve} = y'_{envelope}$. The quantity y'_{curve} is found from equation (12) by differentiating with respect to x for a fixed C:

$$F'_x + F'_y \; y'_{curve} = 0 \qquad (14)$$

Hence, (13) and (14) imply that $F'_C \; C'_x = 0$. But, as it has been indicated, $C (x)$ is a variable quantity and therefore, in general, $C'_x (x) \neq \neq 0$. Consequently, we have

$$F'_C \; (x, y, C) = 0 \qquad (15)$$

Thus, for the points of the envelope, equations (12) and (15) hold simultaneously. Eliminating C from these two equations we arrive at the equation of the envelope.

Example. Let us consider the family of the trajectories of motion of a shell under assumptions enumerated in example 1 of Sec. II.6, when the initial velocity v_0 is given, for different values of the angle of inclination α. Here α serves as a parameter of the family, and therefore, in order to find the envelope (see Fig. 234), we differentiate the equation of the family [equation (II.11)] with respect to α:

$$0 = \frac{x}{\cos^2 \alpha} - \frac{gx^2 \sin \alpha}{v_0^2 \cos^3 \alpha}$$

Expressing $\tan \alpha$ from the last equation and substituting this value into the equation of the family we obtain the equation of the envelope:

$$y = \frac{v_0^2}{2g} - \frac{g}{2v_0^2} \, x^2$$

Hence, the envelope is a parabola, the so-called **safety parabola** (why is it called so?).

Let us take one more example. As we know, all the normals to an evolvent touch the evolute (see Sec. VII.26) and thus the evolute is the envelope of the family of all the normals to the evolvent. This property implies an approximate method of constructing an evolute: we draw several normals to the evolvent and then trace their envelope.

We must take into account that if the curves belonging to a family in question possess singular points (see Sec. 4) then, when elimina-

ting C from (12) and (15), we obtain, besides the envelope, the curve which is the locus of singular points (see Fig. 235). Virtually, as it was shown in Sec. 4, we have $F'_x = F'_y = 0$ for such points and
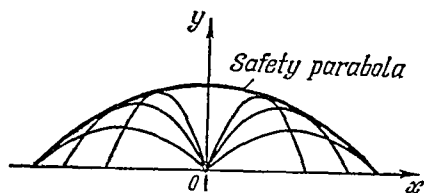
Fig. 234
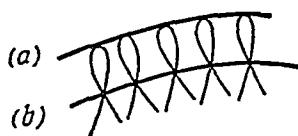
Fig. 235

(a) Envelope (b) Locus of nodal points

therefore in this case (13) implies (15) even if the slopes of the locus of singular points and of the curves of the family do not coincide, that is if $y'_{locus} \neq y'_{curve}$.

§ 2. Extremum of a Function of Several Variables

6. Taylor's Formula for a Function of Several Variables. For definiteness, let us consider a function $f(x, y)$ of two variables. (Similar results are valid for an arbitrary number of independent

Fig. 236

variables.) It turns out that formula (IV.62) remains true for such a function f without any changes.
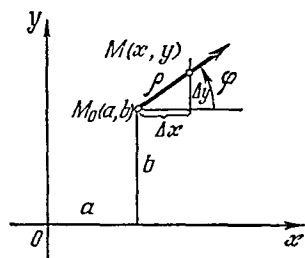
To prove the assertion let us take an arbitrary direction and draw a ray passing through a point (a, b) in the x, y-plane in this direction (which is indicated by the arrow in Fig. 236). The values of the function f which are taken on this ray depend only on the single argument ρ, i.e. $f(x, y) = f^*(\rho)$. We can thus apply formula (IV.62) to the function f^*. We have $\Delta f^* = \Delta f$ but at the same time, in investigating the relationship between the differentials of f^* and f, we must take into account the following fact. According to Fig. 236, we have

$$x = a + \rho \cos \varphi, \quad y = b + \rho \sin \varphi \quad (a, b, \varphi = \text{const}) \quad (16)$$

and therefore

$$f^*(\rho) = f(x, y) = f(a + \rho \cos \varphi, b + \rho \sin \varphi)$$

Hence, in computing df, d^2f, ... we consider the variables x and y to be independent whereas in computing df^*, d^2f^*, ... we

consider them to be dependent on ρ. As we know (see Secs. IX.12 and IX.16), this does not matter for the first-order differentials, that is we have $df^* = df$. but, generally, this is essential for higher-order differentials. But in our case formula (16) implies that

$$d^2x = d^3x = \ldots = 0 \quad \text{and} \quad d^2y = d^3y = \ldots = 0$$

Therefore in this particular case formulas (IX.23) and (IX.24) show that $d^2f^* = d^2f$ and, similarly, $d^3f^* = d^3f$ etc.

Thus, formula (IV.62) for the function f^* automatically implies the validity of the same formula for the function $f(x, y)$.

In practical applications we usually truncate the formula thus retaining only one or two terms. Then we get (see Sec. IX.16) the formulas

$$f(a + h, b + k) = f(a, b) + f'_x(a, b) h + f'_y(a, b) k +$$
$$+ \text{ the terms of the order of smallness not less than the second}$$
$$\text{(relative to } h \text{ and } k) \tag{17}$$

and

$$f(a + h, b + k) = f(a, b) + f'_x(a, b) h + f'_y(a, b) k +$$
$$+ \frac{1}{2} [f''_{xx}(a, b) h^2 + 2f''_{xy}(a, b) hk + f''_{yy}(a, b) k^2] +$$
$$+ \text{ the terms of the order of smallness not less than the third} \tag{18}$$

As in the case of a function of one argument, formulas (17) and (18) can be applied if $|h|$ and $|k|$ are sufficiently small because otherwise the formulas can lead to incorrect results. In all cases when we apply Taylor's formula we suppose that the corresponding derivatives exist and are finite.

7. Extremum. As in Sec. 6. we shall take, for the sake of simplicity, the case of a function $z = f(x, y)$ of two arguments. The definition of an extremum is similar to the one introduced for functions of one independent variable (see Sec. IV.18). For instance, we say that a function $z = f(x, y)$ has a maximum at "a point" (that is for certain values $x = x_0$ and $y = y_0$) if the value $f(x_0, y_0)$ is greater than all the "neighbouring" values of the function f, i.e. than the values $f(x, y)$ taken for x and y which are, respectively, sufficiently close to x_0 and y_0.

In this section we shall consider only extrema that are attained in the interior of the domain of definition of a function f and, besides, we shall suppose that the function f itself and its partial derivatives have no discontinuities. Fig. 237 approximately represents the disposition of the family of level lines of a function f of this type (see Sec. IX.1) near its point of extremum.

We can easily establish the necessary condition for an extremum. Indeed, if we fix $y = y_0$ and make x vary then the corresponding

point (x, y) in Fig. 237 will move along the straight line ll, and the function $f = f(x, y_0)$ (regarded as a function of x) will have an extremum at $x = x_0$. The function $f = f(x, y_0)$ depending only on x, we have $f'_x(x_0, y_0) = 0$, according to Sec. IV.18. This is a partial derivative since it is taken for a fixed y. We similarly consider the case when the point (x, y) moves along the straight line mm and thus deduce the following necessary conditions for an extremum:

$$f'_x(x_0, y_0) = 0, \quad f'_y(x_0, y_0) = 0 \tag{19}$$

(in the case of a function of a greater number of independent variables we must similarly equate to zero all the partial derivatives of the first order). A point (lying in the x, y-plane) at which conditions (19) hold is called a **critical (stationary) point** of the function f. Consequently, if the conditions imposed in the preceding paragraph hold all the points of extremum of the function f are its critical points.
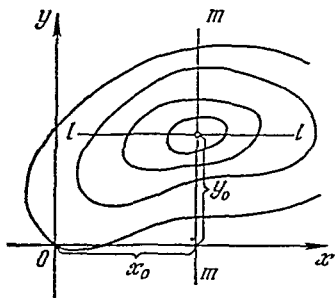
Fig. 237

Conversely, let a critical point (x_0, y_0) of the function $f(x, y)$ be found. Is it correct to assert that this point must be a point of extremum? If there is only one critical point in a certain region and if we are sure that some physical or other conditions guarantee the existence of the extremum then our answer is affirmative. In other cases we must apply some sufficient conditions which we are going to study now.

As we know from Sec. IV.18, the necessary condition for an extremum of a function $f(x)$ of one independent variable which is expressed by the equality $f'(x_0) = 0$ is at the same time "almost sufficient" because if in addition we have $f''(x_0) \neq 0$ the extremum at the point $x = x_0$ is sure to exist. But it turns out that for the case of a function of several variables this is not so. If conditions (19) hold for $f(x, y)$ and the partial derivatives of the second order of the function of two variables are different from zero at the point (x_0, y_0) the extremum nevertheless may not exist. Thus, in the case of functions of many variables we can have a situation of a new type.

For example, see the "graph" of the function $z = f(x, y) = x^2 + y^2$ depicted in Fig. 220. Conditions (19) yield a single stationary point in this case which is the point $(0, 0)$. Evidently, there is a minimum at this point because $f(0, 0) = 0$ and $z > 0$ at the other points. At the same time, for the function $z = -x^2 + y^2$

we have an essentially different case. Its "graph" is shown in Fig. 221. Again, the only critical point is the origin of coordinates. We have $z = y^2$ for $x = 0$, that is the function increases when we move from the origin along the y-axis in both directions and, as a function of one variable y, it has a minimum at the origin. But if $y = 0$ then $z = -x^2$ and hence along the x-axis the function decreases in both directions and has a maximum at the origin. Taking other straight lines passing through the origin we see that the function also has a maximum at the origin for one group of these straight lines and has a minimum for the other group (by the way the function assumes the constant value $z=0$ on two straight lines which are $y = \pm x$). In such a case we say that the function has a **minimax** at the point in question. Hence, the function $z = -x^2 + y^2$ has neither a maximum nor a minimum at the origin although necessary conditions (19) are fulfilled in this case and the partial derivatives $\frac{\partial^2 z}{\partial x^2} = -2$ and $\frac{\partial^2 z}{\partial y^2} = 2$ are different from zero.

After discussing these examples let us proceed to investigate the functions of general form. Suppose conditions (19) are fulfilled at a point (x_0, y_0) for a function $f(x, y)$. Let us see whether or not the function f has an extremum at the point. In order to do this we take Taylor's formula (18) and put $a = x_0$, $b = y_0$ in it. This results in

$$\Delta f = f(x_0 + h, y_0 + k) - f(x_0, y_0) =$$
$$= \frac{1}{2} [f''_{xx}(x_0, y_0) h^2 + 2f''_{xy}(x_0, y_0) hk + f''_{yy}(x_0, y_0) k^2] +$$
$$+ \textit{the terms of the order of smallness not less than the third}$$

The terms of the first order of smallness relative to h and k do not enter in the result since the stationarity conditions (19) imply that these terms are equal to zero. The terms of the third order being considerably smaller than the terms of the second order for sufficiently small $|h|$ and $|k|$, the sign of the right-hand side is determined by the group of terms of the second order. Hence the sign coincides with the sign of the quadratic form

$$P(h, k) = f''_{xx}(x_0, y_0) h^2 + 2f''_{xy}(x_0, y_0) hk + f''_{yy}(x_0, y_0) k^2 \quad (20)$$

(we do not put down the factor $\frac{1}{2}$ here since it does not affect the sign we are interested in). Consequently, if the sum (20) is positive for all h and k (of course, except for $h = k = 0$ when it vanishes) then we have $\Delta f > 0$ for sufficiently small $|h|$, $|k|$ which implies that $f(x_0 + h, y_0 + k) > f(x_0, y_0)$. Hence, we have a minimum at the point (x_0, y_0) in this case. If the sum is negative then we likewise conclude that there will be a maximum at the point (x_0, y_0). Finally, if the sum can assume the values of both signs there will

be a minimax at the point (x_0, y_0) and thus there will be no extremum. We cannot judge by the sum of the terms of the second order whether there is an extremum only when the sum can vanish for certain values of h and k which are not equal to zero but does not change its sign (in particular, this is the case when the sum is identically equal to zero and does not enter into the expression of Δf). In such a case we must take into account the subsequent terms of Taylor's formula and take the sum of the terms of the third order. Then a similar (but, of course, more complicated) investigation of the sum for the values of h and k which turn the sum of the terms of the second order into zero can indicate whether we have an extremum and so on. We are not going to carry out this investigation here.

These conclusions can be similarly drawn for functions of any number of independent variables. But in the case of a function of two variables we can easily proceed to express the sufficient conditions for the extremum directly in terms of the values of the second derivatives at the point (x_0, y_0). To do this let us take h^2 outside the brackets on the right-hand side of (20) and denote $\frac{k}{h} = t$. Then we obtain

$$P\,(h,\ k) = [(f''_{xx})_0 + 2\,(f''_{xy})_0\,t + (f''_{yy})_0\,t^2]\,h^2 \qquad (21)$$

[the subscript "zero" indicates that the values of the derivatives are taken at the stationary point (x_0, y_0)]. As is well known from elementary algebra, the polynomial in t inside the square brackets has two distinct real roots if its discriminant is positive, i.e. if

$$(f''_{xy})_0^2 - (f''_{xx})_0\,(f''_{yy})_0 > 0 \qquad (22)$$

In this case the polynomial changes its sign when passing through the roots, and we therefore have a minimax here. But if

$$(f''_{xy})_0^2 - (f''_{xx})_0\,(f''_{yy})_0 < 0$$

then the polynomial has imaginary roots and consequently it does not change its sign (why is it so?). Therefore we have an extremum in this case. To find out what is the sign of the right-hand side of (21) we put $t = 0$. Then we see that if

$$(f''_{xy})_0^2 - (f''^1_{xx})_0\,(f''_{yy})_0 < 0, \qquad (f''_{xx})_0 > 0 \qquad (23)$$

then the right-hand side of (21) is positive for all t and thus, by the results of the preceding paragraph, the function f has a minimum at the point (x_0, y_0). Similarly, if

$$(f''_{xy})_0^2 - (f''_{xx})_0\,(f''_{yy})_0 < 0, \qquad (f''_{xx})_0 < 0 \qquad (24)$$

then the function f has a maximum. Finally, if

$$(f''_{xy})_0^2 - (f''_{xx})_0\,(f''_{yy})_0 = 0 \qquad (25)$$

then the polynomial entering into (21) has a double root and thus it does not change its sign but it can vanish for some nonzero values of h and k. Thus, this is an ambiguous case.

The condition guaranteeing that quadratic form (20) should be positive can also be deduced from the general theory of quadratic forms (see Sec. XI.11). According to the theory, after a certain rotation of the coordinate axes, form (20) has been transformed to a "diagonal form", that is to the form

$$P = \lambda_1 h'^2 + \lambda_2 k'^2 \tag{26}$$

where λ_1 and λ_2 are the roots of the characteristic equation

$$\begin{vmatrix} (f''_{xx})_0 - \lambda & (f''_{xy})_0 \\ (f''_{xy})_0 & (f''_{yy})_0 - \lambda \end{vmatrix} = 0 \tag{27}$$

and h' and k' are the increments of the new coordinates (which occur after the rotation has been performed). Equation (27) implies that

$$\lambda_1 \lambda_2 = (f''_{xx})_0 (f''_{yy})_0 - (f''_{xy})_0^2, \qquad \lambda_1 + \lambda_2 = (f''_{xx})_0 + (f''_{yy})_0$$

(check it up!). We can easily deduce from these equalities that in cases (22)-(25) we respectively have $\lambda_1 \lambda_2 < 0$ or $\lambda_1 > 0$, $\lambda_2 > 0$ or $\lambda_1 < 0$, $\lambda_2 < 0$ or $\lambda_1 \lambda_2 = 0$ (we leave it to the reader to verify these assertions). From this, on the basis of equality (26), we deduce the same conditions as those in the preceding paragraph.

In order to investigate the behaviour of a function $f(x_1, x_2, \dots \dots, x_n)$ depending on an arbitrary number n of arguments at a stationary point we must take the quadratic form

$$\sum_{i,\,j=1}^{n} (f''_{x_i x_j})_0 h_i h_j \tag{28}$$

instead of (20) and the equation

$$\det (A - \lambda I) = 0 \tag{29}$$

[where $A = ((f''_{x_i x_j})_0)_{nn}$] in place of (27). If all the roots of equation (29) are positive then form (28) assumes only positive values at all points $(h_1, h_2, \dots, h_n)$ different from the point $h_1 = h_2 = \dots = h_n = 0$. A quadratic form of this type is said to be **positive definite**. In this case the function f has a minimum at the stationary point in question. If all the roots of equation (29) are negative then form (28) is **negative definite**, and the function f has a maximum. But if equation (29) possesses roots of both signs then the function f has a minimax.*

---

* It can be shown that an equation of form (29) with the matrix $A = ((f''_{x_i x_j})_0)$ has only real roots because the matrix is symmetric. See also Sec. XI.10 on this question.—*Tr.*

8. **The Method of Least Squares.** As an example illustrating applications of the theory of extremum of functions of several arguments we shall consider the least-square method which can be applied to constructing empirical formulas. In particular, the method is used when the accuracy of the approximate method described in Sec. I.30 is insufficient. It is also applied to automatizing calculations. Here we shall dwell only on the problem of selecting a linear functional relationship in the case of one independent variable. In performing these calculations we usually apply the following way of reasoning which can also be applied to other functional relationships: the sought-for function is of the form $y = kx + b$ but the values of the parameters k and b are yet unknown. The substitution of $x = x_i$ into the formula should have resulted in $kx_i + b$ but the experimental data give the value y_i, and thus we have the difference $y_i - kx_i - b$ between the theoretical and experimental data which is due to the errors of the experiment and of the calculations, the non-linearity of the relationship under consideration etc. This difference between the left-hand side and the right-hand side of a formula is called a **discrepancy**.

Therefore, let us try to select k and b in such a way that the sum of the squares of these discrepancies, that is the quantity

$$S = \sum_{i=1}^{N} (y_i - kx_i - b)^2$$

should take on its minimal value among all the possible values. We can also take a sum of other even powers or, for instance, the sum of the absolute values of the discrepancies, but this will involve more complicated calculations. At the same time we must not take the sum of the discrepancies themselves because it can be small when the absolute values of the summands (which can be of different signs) are large. Thus we arrive at the problem of finding a minimum of the function $S = S(k, b)$. Applying necessary conditions (19) we see that for the function to have a minimum, it is necessary that the equalities

$$S'_k = - \sum_{i=1}^{N} 2(y_i - kx_i - b)x_i = 0, \quad S'_b = - \sum_{i=1}^{N} 2(y_i - kx_i - b) = 0$$

should be fulfilled. It follows that

$$k \sum_{i=1}^{N} x_i^2 + b \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i, \quad k \sum_{i=1}^{N} x_i + bN = \sum_{i=1}^{N} y_i$$

Hence, we have obtained a simple system of two algebraic equations of the first degree in two unknowns from which k and b can be found. All x_i and y_i being known, the system can easily be solved. The fact that in this way we really obtain a minimum of S is implied by the meaning of the problem in question.

Similar methods are applied to selecting other empirical formulas and to some other problems.

Let us take an example. Suppose that an experiment carried out in order to establish a relationship between some quantities x and y indicates that these relationships are approximately expressed by the equations

$$\left.\begin{array}{r} x + y = 5.8 \\ x + 2y = 8.1 \\ 2x + 3y = 13.2 \end{array}\right\} \qquad (30)$$

The system is formally inconsistent because adding together the first two equations we arrive at a contradiction with the third equation. But this can be a result of the errors of the experiment! Therefore let us try to satisfy system (30) as precisely as possible so that the sum of the squares of the discrepancies should be as small as possible. Thus, we are going to find the values of x and y such that the quantity

$$S = (x + y - 5.8)^2 + (x + 2y - 8.1)^2 + (2x + 3y - 13.2)^2$$

assumes its minimal value. Applying necessary conditions (19) we obtain

$$S'_x = 2\ (x+y-5.8)+2\ (x+2y-8.1)+2\ (2x+3y-13.2)\ 2=0,$$
$$S'_y = 2\ (x+y-5.8)+2\ (x+2y-8.1)\ 2+2\ (2x+3y-13.2)\ 3=0$$

Cancelling out the factor 2 we get

$$\left.\begin{array}{r} 6x + 9y = 5.8 + 8.1 + 2 \times 13.2 = 40.3 \\ 9x + 14y = 5.8 + 2 \times 8.1 + 3 \times 13.2 = 61.6 \end{array}\right\}$$

Solving the equations in the simplest way we find $x = 3.3$ and $y = 2.3$. Of course, these values only approximately satisfy system (30). Naturally, the greater the number of relationships of form (30) between x and y, the more reliable the values of x and y thus obtained. Of course, this is so if there are no systematic errors in the experiment (random errors occurring in some relationships mutually cancel). Many other systems of approximate equations can be solved in like manner. In particular, the method can be applied to systems of empirical equations when the number of equations exceeds the number of unknowns.

The method of least squares was discovered by the French mathematician A. M. Legendre (1752-1833) and by Gauss. It has many useful applications at present time.

9. Curvature of Surfaces. The classification of stationary points described in Sec. 7 is directly related to the classification of surfaces.

Let us consider an arbitrary surface (S) and take a point M on it (see Fig. 238). If we draw the normal nn to the surface at the point and then draw an arbitrary plane (P) passing through the normal the plane will intersect (S) along a plane curve ll which is called a normal section of (S) at M. The curve ll has a certain curvature k

at the point M (see Sec. VII.24). Now if we rotate the plane (P) about the normal nn the normal section will vary, and its curvature k will therefore also vary, in the general case. To investigate the
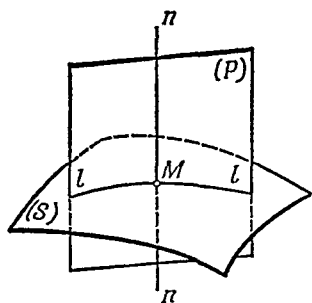
Fig. 238

law of this variation let us choose a Cartesian coordinate system so that the origin of the coordinates should be at the point M and the z-axis should go along the normal nn. Then in the vicinity of M the surface (S) can be represented by an equation of the form $z = z(x, y)$, and the point M [at which $z(0, 0) = 0$] will be a stationary (critical) point of the function $z(x, y)$ (why will it be so?). By arguments similar to the ones applied at the end of Sec. 7 we conclude that after a certain rotation of the coordinate axes about the axis nn has been performed the equation of (S) turns into the form

$$z = \frac{1}{2}(\lambda_1 x'^2 + \lambda_2 y'^2) + \textit{the terms of higher order of smallness} \quad (31)$$

where x' and y' are the new coordinates replacing x and y. Let the plane (P) form an angle φ with the plane $x'Mz$. Then passing to polar coordinates we obtain $x' = \rho \cos \varphi$ and $y' = \rho \sin \varphi$ which yield

$$z = \frac{1}{2}(\lambda_1 \cos^2 \varphi + \lambda_2 \sin^2 \varphi) \rho^2 + \ldots$$

Hence, at the point M we have $\frac{dz}{d\rho} = 0$ and $\frac{d^2 z}{d\rho^2} = \lambda_1 \cos^2 \varphi + \lambda_2 \sin^2 \varphi$. Formula (VII.37) thus implies the expression $k = |\lambda_1 \cos^2 \varphi + \lambda_2 \sin^2 \varphi|$ for the curvature. Accordingly, there can be three cases here, namely the following cases:

1. Let $\lambda_1 \lambda_2 > 0$, i.e. let λ_1 and λ_2 be of the same sign. Then all the normal sections have the same direction of convexity near the point M, and the values of k lie within the limits $|\lambda_1|$ and $|\lambda_2|$. Besides, we have $k = |\lambda_1|$ for $\varphi = 0$, that is for the plane $x'Mz$, and $k = |\lambda_2|$ for $\varphi = \frac{\pi}{2}$, that is for the plane $y'Mz$ (these are the so-called **principal normal sections**). A point M of this type is called an **elliptic** (or **umbilical** in case $|\lambda_1| = |\lambda_2|$) **point** of the surface (S). For instance, all the points of an ellipsoid or of a hyperboloid of two sheets are elliptic. Equation (31) implies that the tangent plane to (S) at the point M has only one common point M with the surface (S) near the point M. The planes parallel to the tangent plane which are drawn sufficiently close to it and which intersect (S) yield the intersection lines of the form of an infinitesi-

mal ellipse whose axes lie in the planes of principal normal sections. The form of the sections can become more complicated when the distance from the plane (parallel to the tangent plane) to the point M is increased (see Fig. 239a).

2. Let $\lambda_1\lambda_2 < 0$. Then some of the normal sections have a positive curvature at the point M and the direction of convexity coinciding with the direction of the outer normal to the surface near the point M, and some other sections have a negative curvature and the opposite direction of convexity. For instance, this is the

(a)   (b)   (c)
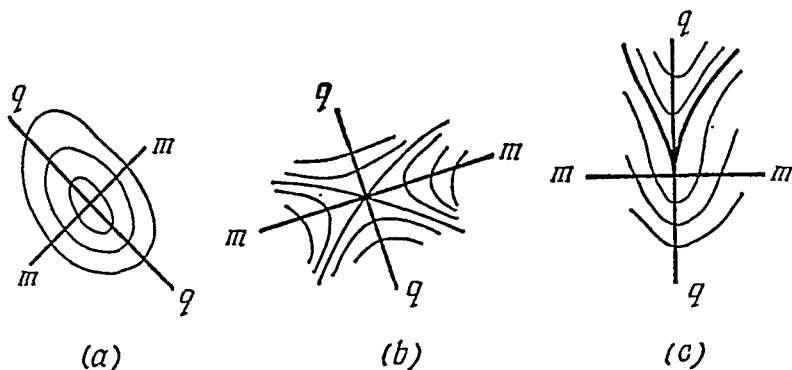
Fig. 239
mm and qq are principal normal sections

case for the points of a hyperboloid of one sheet. One of the sections belonging to the first group has the maximal curvature $|\lambda_1|$ (or $|\lambda_2|$) and one of the sections of the second group has the maximal curvature $|\lambda_2|$ (or, respectively, $|\lambda_1|$). These principal normal sections are also mutually perpendicular. A point M of this type is called a **hyperbolic point** of the surface. The tangent plane to (S) at the point M intersects the surface (S) along two curves which form a nonzero angle (i.e. the angle between the tangent lines to the curves) at M. The planes parallel to the tangent plane which are drawn infinitely close to M intersect (S) along hyperbolas lying infinitely close to M, the axes of the hyperbolas being directed along the principal normal sections (see Fig. 239b).

3. Let $\lambda_1\lambda_2 = 0$. Then, if λ_1 and λ_2 are not simultaneously equal to zero all the normal sections have the same direction of convexity near M and have a nonzero curvature at M except for one of the sections which has the zero curvature at M. The curvature of the section which is perpendicular to the section of zero curvature is the maximal at the point M. A point of this type is called a **parabolic point**. For example, all the points of a cylindrical or conical surface are of this type. A typical disposition of lines of intersections of a surface (S) with the planes parallel to the tangent plane

drawn at a parabolic point M of the surface is shown in Fig. 239c but there can be some other dispositions in different cases. The case when $\lambda_1 = \lambda_2 = 0$, that is when $k = 0$ for all φ, also belongs to this type. In such a case a point of this type is called a **planar point** of the surface (S). It is clear that all the points of a plane are of this type.

In the above concrete examples all the points of each of the surfaces were of the same type but this must not be necessarily so in the general case. For instance, the surface of a torus has points belonging to each of the three types (where are these points placed on the surface of a torus?).

In all cases the product $\lambda_1 \lambda_2$ is called the **total (Gaussian) curvature of the surface** (S) **at the point** M. There is a remarkable property of the total curvature: when a surface is bent without stretching its total curvature does not change. For instance, if we take a sheet of paper and bend it in an arbitrary way the surface thus obtained will have the zero total curvature at each of its points. The same cause makes it impossible to flatten a portion of a sphere without deformation, and there cannot therefore be a geographic map without distortion.

A surface can have **singular points** (usually these are **isolated points** or **"conical" points** similar to the vertex of a circular cone but of course there are also singular points of more complicated types). It can also have **singular lines** which are loci of singular points (most often these lines are isolated lines or lines of self-intersection; there are also **"cuspidal edges"** and some other types of singular lines).

10. Conditional Extremum. In the problems considered in Sec. 7 we investigated extrema in the case when independent variables were not connected by any additional relationships. An extremum of this kind is called an **unconditional extremum**. But there are also problems concerning a so-called **conditional extremum** when arguments are related to one another by relationships of the form of an equality. We begin our investigation with functions of two independent variables.

Suppose we seek for a maximum or a minimum of a function $z = f(x, y)$ on the condition that x and y are restricted by the relationship

$$F(x, y) = h \tag{32}$$

Equation (32) is called a **coupling equation** [conditions of form (32) are also called **subsidiary conditions, side conditions** or **constraints**]. Thus, we consider and compare only those values of the function f that correspond to the points (lying in the x, y-plane) which belong to the curve represented by equation (32). For instance, in Fig. 240 we see level lines of a function f. At the point K the function attains

its maximum (unconditional). At the same time there are three conditional extrema here, namely two maxima at the points A and C and one minimum at the point B (think why it is so). An unconditional maximum can be compared to the top of a mountain. Then it is natural to compare a conditional maximum to the highest point of a mountain path whose projection on the x, y-plane has an equation of form (32).

If it is possible to express y in terms of x with the aid of equation (32) then we can substitute the result $y = y(x)$ into the expression
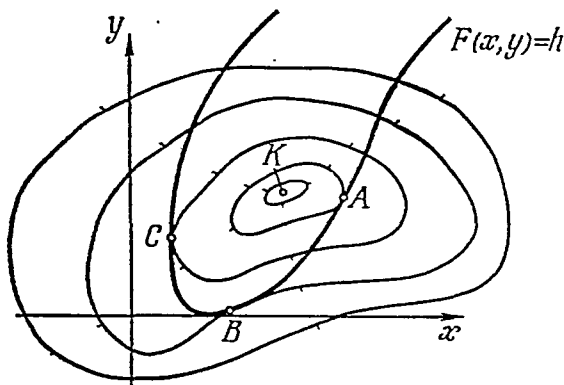
Fig. 240

of z and thus obtain z as a function of a single independent variable:

$$z = f[x, y(x)] \tag{33}$$

When substituting $y = y(x)$ we have taken into account condition (32), and, since there are no other restrictions, the problem reduces to finding an unconditional extremum of $z = f[x, y(x)]$. There will be a similar situation if it is possible to solve equation (32) for x or if the curve defined by equation (32) can be represented by parametric equations.

It should be noted that the above resolution of equation (32) may not be possible in some cases and, besides, this can be inconvenient and can lead to complicated expressions even when equation (32) is solvable. In such a case we can reason in the following way. Coupling equation (32) defines some relationship $y = y(x)$ which may not be known in an explicit form. Consequently, z is a composite function of the independent variable x of form (33). The necessary condition for an extremum can therefore be put down in the form

$$\frac{dz}{dx} = f'_x + f'_y \frac{dy}{dx} = 0 \tag{34}$$

on the basis of the formula for the derivative of a composite function. Here the expression $\frac{dy}{dx}$ designates the derivative of the implicit function $y\,(x)$ defined by condition (32). Hence, by Sec. IX.13, we have

$$F'_x + F'_y \frac{dy}{dx} = 0, \quad \text{i.e.} \quad \frac{dy}{dx} = -\frac{F'_x}{F'_y}$$

Substituting this expression into (34) we see that at the point of a conditional extremum we have

$$f'_x - \frac{F'_x}{F'_y} f'_y = 0, \quad \text{i.e.} \quad \frac{f'_x}{F'_x} = \frac{f'_y}{F'_y}$$

Let us denote the value of the last ratio by λ. Then we have

$$\frac{f'_x}{F'_x} = \frac{f'_y}{F'_y} = \lambda \tag{35}$$

at the point of a conditional extremum. This can be rewritten as

$$f'_x - \lambda F'_x = 0, \quad f'_y - \lambda F'_y = 0 \tag{36}$$

Let us introduce the notation

$$f^* \,(x,\, y,\, \lambda) = f\,(x,\, y) - \lambda F\,(x,\, y) \tag{37}$$

where λ is an undetermined parameter which is called **Lagrange's undetermined multiplier (factor)** named after Lagrange who introduced this method. Then equations (36) can be put down in the form

$$f^{*\prime}_x = 0, \quad f^{*\prime}_y = 0 \tag{38}$$

Thus, we have arrived at equations of the same form [see equations (19)] but for the modified function f^* defined by formula (37) instead of the original function f. Equations (38) together with the coupling equation (32) form a system of three equations in three unknowns which are x, y and λ. The points in the x, y-plane at which a conditional extremum may be attained are found from these equations. The conditions thus obtained are only necessary. Sufficient conditions guaranteeing the existence of a conditional extremum at a point defined by equations (38) and (32) can similarly be deduced from the sufficient conditions for an unconditional extremum established above but we shall not do this here. [By the way, in our case it is sufficient to compute the derivative $\frac{d^2z}{dx^2}$ and to investigate its sign at the point defined by (32) and (38).]

Lagrange's multiplier λ has a simple meaning. To illustrate it denote the coordinates of the point of a conditional extremum by $\bar{x}$ and $\bar{y}$. Let $\bar{z}$ designate the corresponding extremal value of z. Up till now the quantity h entering into equation (32) has been considered to be fixed. But now we can vary h and then all the

three quantities $\bar{x}$, $\bar{y}$ and $\bar{z}$ will also vary, i.e. $\bar{x}$, $\bar{y}$ and $\bar{z}$ will become functions of h. The identity $\bar{z}(h) \equiv f[\bar{x}(h), \bar{y}(h)]$ holding, we have

$$\frac{d\bar{z}}{dh} = f'_x \frac{d\bar{x}}{dh} + f'_y \frac{d\bar{y}}{dh} \tag{39}$$

On the other hand, by (32), we have

$$F'_x \frac{d\bar{x}}{dh} + F'_y \frac{d\bar{y}}{dh} = 1 \tag{40}$$

From (39), (35) and (40) we readily deduce $\frac{d\bar{z}}{dh} = \lambda$. Hence, the factor λ is equal to the rate of change of the extremal value $\bar{z}$ when the parameter h entering into coupling equation (32) varies.

The investigation of a conditional extremum in the general case of an arbitrary number of independent variables and any number of coupling equations is carried out in like manner. [We remind the reader that according to Sec. X.2 the number of coupling equations (subsidiary conditions) must be less than the number of arguments.]

For instance, if we are looking for an extremum of a function $f(x, y, z, u, v)$ when the arguments x, y, z, u and v are restricted by the conditions

$$F_1(x, y, z, u, v) = 0, \quad F_2(x, y, z, u, v) = 0 \quad \text{and}$$
$$F_3(x, y, z, u, v) = 0 \tag{41}$$

we must perform calculations as if we wanted to find an unconditional extremum of the function

$$f^* = f - \lambda_1 F_1 - \lambda_2 F_2 - \lambda_3 F_3$$

where λ_1, λ_2 and λ_3 are undetermined Lagrange's multipliers. The stationarity condition for f^* yields the equations

$$f^{*\prime}_x = 0, \quad f^{*\prime}_y = 0, \quad f^{*\prime}_z = 0, \quad f^{*\prime}_u = 0 \quad \text{and} \quad f^{*\prime}_v = 0$$

which, together with equations (41), form a system of $8 = 5 + 3$ equations in $8 = 5 + 3$ unknowns x, y, z, u, v, λ_1, λ_2 and λ_3.

Methods of solving a problem of a conditional extremum form the basis of one of the widely spread numerical methods of finding an unconditional extremum. This is the so-called **method of steepest descent**. Here we shall describe a variant of the method applicable to the case of a minimum of a function of two arguments although in the general case different modifications of the method can be applied to finding minima (or maxima) of functions of any number of independent variables. Let it be necessary to find a point of

25*

minimum of a function $f(x, y)$. It turns out that when the function f does not belong to a certain class of the simplest functions it is rather difficult (and even impossible) to solve equations (19). Besides, in solving equations (19) we find all the stationary points including those which are not points of minimum and hence we do much unnecessary work. It is therefore better to apply the method of steepest descent which is an iterative method. We begin with taking a point $M_0 (x_0, y_0)$ as a zeroth approximation. At this point the function f has the maximal rate of decrease in the direction of the vector $- (\text{grad } f)_{M_0} = -f'_x (x_0, y_0) \, \mathbf{i} - f'_y (x_0, y_0) \, \mathbf{j}$ (because, as was shown in Sec. 1, the direction of the vector grad f indicates the direction of the maximal rate of increase of the function). Let us draw a ray through the point M_0 in this direction and consider the values of the function f which are taken on the ray. These values are expressed by the quantity $f(x_0 - f'_{x_0} t, \, y_0 - f'_{y_0} t)$ which we regard as a function of t for $t > 0$. After that we find a value of t for which this function of one variable attains its minimum. This value of t determines a new point $M_1 (x_1, y_1)$. Then we draw a ray through the point M_1 in the direction of the vector $-(\text{grad } f)_{M_1}$ [$(\text{grad } f)_{M_1}$ designates the gradient at the point M_1] and find a point M_2 at which f attains its minimum on the ray etc. In many cases this method enables us to find an approximate position of the sought-for point of extremum within a sufficient accuracy after several steps of this kind have been carried out. (We suggest that the reader should consider level lines of a function f on the x, y-plane and find out the geometric meaning of the method.) Methods of this type which enable us to find extrema without using necessary conditions are called direct methods.

11. Extremum with Unilateral Constraints. Independent variables involved in an extremal problem can be restricted by one or several conditions of the form of an inequality. Such conditions are called unilateral (one-sided or non-restricting) constraints. For example, let an extremum of a function $f(x, y)$ be sought for and let the independent variables be related to one another by the constraint $F(x, y) \geqslant 0$ which defines a domain (S) with a boundary (L) in the x, y-plane (see Fig. 241). The curve (L) is defined by the equation $F = 0$. The function f can have both extrema attained in the interior of (S) and extrema attained on (L). In order to find the former we can use conditions (19) defining stationary points but these conditions do not apply to extrema on the boundary (L). To find the latter extrema we remark that if the function f has an extremum at a point M belonging to (L), for instance, a minimum, then the value $f(M)$ is smaller than all the values of f taken on (L) near M. Therefore there will simultaneously be a conditional minimum of f at M for the coupling equation $F = 0$. Hence, such points can be found with the help of the methods described in Sec. 10.

These results can be applied to the problem of finding the greatest and the least values of a function. For instance, let a function $z = f(x, y)$ be considered in the domain depicted in Fig. 242. Suppose that neither the function nor its derivatives have discontinuities in the domain. If the point at which the function attains its greatest value lies in the interior of the domain then there will be an unconditional maximum of the function at this point. If the point belongs to the contour bounding the domain but does not lie at the vertices A, B and C then there will be a conditional maximum of the function at this point, and the equation of the corresponding arc of the contour will serve as condition (32). Finally, if the greatest value
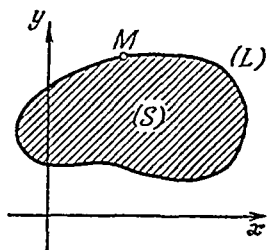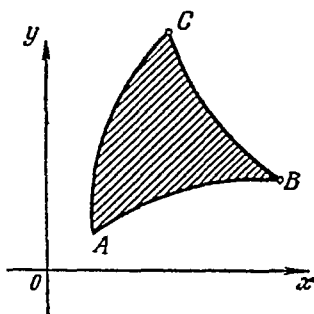
Fig. 241   Fig. 242

is attained at a vertex (A, B or C) then in order to find this greatest value we must additionally compare the values of the function at the points A, B and C with its other extremal values.

Thus, to find the greatest value we must find all the points of unconditional maximum lying in the interior of the domain and all the points of conditional maximum belonging to the contour. Moreover, in case it is difficult to specify beforehand which of the possible points of conditional extremum will be the points of maximum it is advisable to find the values of the function at all these points. After that we compare with each other the extremal (maximal) values taken in the interior of the domain, the extremal values attained on the contour and the values of the function at the points A, B and C and thus find the sought-for greatest value of the function. The least value of a function is found in like manner. As in Sec. IV.19, it is better to seek the greatest and the least values of a function simultaneously.

If the domain under consideration contains points of discontinuity of the partial derivatives of the first order the values of the function at these points should be included into the set of values which we compare with each other because it can turn out that the greatest (or the least) value of the function is attained at some of these points. If there are points of discontinuity of the function then we

should additionally investigate the behaviour of the function in approaching such points. If there are lines of discontinuity of the function or of its derivatives then the values of the function taken on such lines must also be investigated which leads to the problem of a conditional extremum. Finally, if the domain in which we investigate the function extends to infinity we must additionally investigate the behaviour of the function when the variable point in the x, y-plane approaches infinity.

Functions of a greater number of independent variables are investigated in a similar way. But in performing such an investigation

Fig. 243

we must take into account some additional factors which are due to the increase of the dimension. For instance, if we consider a function $u = = f(x, y, z)$ in the domain shown in Fig. 243 then we must consider its unconditional maxima in the interior of the "curvilinear tetrahedron", its conditional maxima with one coupling equation on the "faces" of the tetrahedron (in this case the equations of the corresponding surfaces serve as coupling equations), the conditional maxima with two coupling equations on the "edges" of the tetrahedron [the role of the coupling equations can be played by the equations of the corresponding space curve if they are put down in form (X.2)] and, finally, the values of the function at the "vertices". Comparing all these values with each other we find the greatest value of the function. Similarly, if there are surfaces of discontinuity of the function we come to the problem of a conditional maximum with one coupling equation, and if there are lines of discontinuity we come to the problem of a conditional extremum with two coupling equations etc.
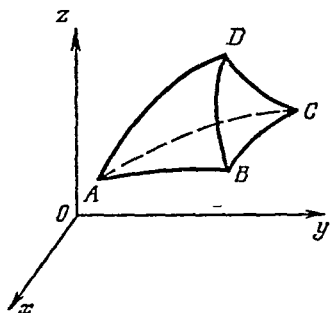
If we apply an iterative scheme of the type of the method of steepest descent described in Sec. 10 then in the case when there are several minima (maxima) in the domain in question we can come to a minimum (maximum) which is not the least (greatest). In such a case it is advisable to apply the method several times beginning with different zeroth approximations chosen at random. For instance, after such repeated calculations have been carried out we can obtain a smaller minimum than the one found in the first calculation, and in many problems this may finally lead to the least value of the function.

12. Numerical Solution of Systems of Equations. In conclusion we shall consider some methods of numerical solution of a system

of two equations in two unknowns. The case of a system of n equations in n unknowns can be treated similarly.

The *iterative method* has the same form as in Secs. V.3 and VI.5. To apply the method we rewrite the system in question in the form

$$\left. \begin{array}{l} x = f(x, y) \\ y = g(x, y) \end{array} \right\} \tag{42}$$

Then we choose a zeroth approximation $x = x_0$, $y = y_0$. The subsequent approximations are constructed according to the formulas

$$\left. \begin{array}{l} x_1 = f(x_0, y_0) \\ y_1 = g(x_0, y_0) \end{array} \right\}, \quad \left. \begin{array}{l} x_2 = f(x_1, y_1) \\ y_2 = g(x_1, y_1) \end{array} \right\} \quad \text{etc.}$$

If the process is convergent then in the limit we obtain a solution of system (42). The smaller the rate of change of the functions f and g when their arguments vary (that is the smaller the absolute values of the partial derivatives of the functions), the better the convergence of the process.

A modification of the method (the **Seidel method**) which is based on using some numerical values obtained at each step of the calculations for computing other values at the same step may sometimes accelerate the convergence. Such calculations are performed according to the following scheme: $x_1 = f(x_0, y_0)$, $y_1 = g(x_1, y_0)$, $x_2 = f(x_1, y_1)$, $y_2 = g(x_2, y_1)$ and so on.

Newton's method (see Sec. V.2) is based on the replacement of given functions by their linear approximations constructed with the help of the values of the functions and their derivatives for the values of the arguments approximately equal to the sought-for solutions. Suppose we have to solve a system of equations of the form

$$\left. \begin{array}{l} P(x, y) = 0 \\ Q(x, y) = 0 \end{array} \right\} \tag{43}$$

Let us begin with a zeroth approximation $x = x_0$, $y = y_0$ to the sought-for solution which can be found by means of an approximate sketch of curves (43) in the x, y-plane or which can be implied by the physical meaning of the problem and the like. Taking expansions (17) of the functions P and Q into powers of $h = x - x_0$ and $k = y - y_0$ and dropping the terms of higher order of smallness we arrive at the following system of equations:

$$\left. \begin{array}{l} P(x_0, y_0) + P'_x(x_0, y_0)(x - x_0) + P'_y(x_0, y_0)(y - y_0) = 0 \\ Q(x_0, y_0) + Q'_x(x_0, y_0)(x - x_0) + Q'_y(x_0, y_0)(y - y_0) = 0 \end{array} \right\} \tag{44}$$

System (44) approximately replaces system (43). Solving (44) which is a system of linear equations we obtain the values of the first approximation $x = x_1$ and $y = y_1$. The second approximation is

found from system (44) after x_1 and y_1 are substituted for x_0 and y_0 in it and so on. The relationship between the nth and $(n + 1)$th approximations is of the form

$$P\,(x_n,\,y_n)+P'_x\,(x_n,\,y_n)\,(x_{n+1}-x_n)+P'_y\,(x_n,\,y_n)\,(y_{n+1}-y_n)=0 \Big\}$$
$$Q\,(x_n,\,y_n)+Q'_x\,(x_n,\,y_n)\,(x_{n+1}-x_n)+Q'_y\,(x_n,\,y_n)\,(y_{n+1}-y_n)=0 \Big\}$$

If the process is convergent we pass to the limit as $n \to \infty$. In the limit the last two summands in each of the equations vanish and thus we see that the limiting values satisfy system (43). It can be shown that if the initial approximation is chosen sufficiently close to the sought-for solution and if the Jacobian (see Sec. IX.13) $\dfrac{D\,(P,\,Q)}{D\,(x,\,y)}$ is unequal to zero, i.e. if

$$\frac{D\,(P,\,Q)}{D\,(x,\,y)} \neq 0$$

then the approximations are sure to converge. [How is the Jacobian related to system (44)?]

We can also take advantage of the fact that a solution of system (43) simultaneously gives minimum to the function $V\,(x,\,y) = [P\,(x,\,y)]^2 + [Q\,(x,\,y)]^2$ (why is it so?). Instead of this function we sometimes take a similar expression in which there are certain numerical positive coefficients in front of the squares of the functions which are introduced to balance the "significance" of both equations (43). Taking a function V of this type we then find its minimum with the help of one of the direct methods mentioned in Secs. 10 and 11. [Obviously, it is senseless to apply the necessary conditions for an extremum to the function V because this will lead us back to system (43). Let the reader verify this assertion.] If the minimal value thus found is equal to zero the point of the minimum yields a solution of system (43).

Indefinite Integral

§ 1. Elementary Methods of Integration

1. Basic Definitions. Let the function $f(x)$ be the derivative of a function $F(x)$, i.e. $F'(x) = f(x)$. Then $F(x)$ is said to be an **antiderivative** (or **primitive**) of $f(x)$. For instance, the function $3x^2$ is the derivative of x^3, and x^3 is an antiderivative of $3x^2$.

Differential calculus deals with the basic problem of finding the derivative of a given function and with the problem of finding its differential which is directly related to the former problem. For functions of one independent variable, this problem was considered in Chapter IV. In particular, as it was shown in Sec. IV.5, the derivative of any elementary function is an elementary function which is found by means of standard rules.

The main problem of integral calculus is reverse to that of differential calculus. This is the problem of finding a function when the derivative of the function is given, that is the problem of finding antiderivatives of a given function. The significance of the problem will be discussed in Chapter XIV. This problem is more complicated than the problem of differentiation. (As a rule, "reverse" problems are more complicated than the "direct" ones. For instance, the problem of extracting a root is more complicated than the problem of raising to a power.) In particular, we shall see that although the antiderivative of any elementary function exists (and these are the functions that we shall deal with in the present chapter) it may not be an elementary function.

A given function possesses more than one antiderivative. For example, we have not only $(x^3)' = 3x^2$ but $(x^3 + 5)' = 3x^2$ as well. (It often turns out, in different examples, that solutions of reverse problems are not unique.) In general, if a function $f(x)$ has antiderivatives $F_1(x)$ and $F_2(x)$ then $F_1' = f$ and $F_2' = f$, i.e. $F_1' - F_2' = 0$, $(F_1 - F_2)' = 0$, and thus we have $F_1 - F_2 = \text{const}$ (see Sec. IV.17) and $F_1 = F_2 + \text{const}$. Consequently, *any two antiderivatives of the same function differ in a constant summand.* Hence, in order to obtain all the antiderivatives of a given function it is.

sufficient to take one of the antiderivatives and to add an arbitrary constant to it. For instance, the family of all antiderivatives of the function $3x^2$ is given by the formula $x^3 + C$ where C is an arbitrary constant. Making C assume concrete numerical values we obtain particular antiderivatives: x^3, $x^3 + 5$, $x^3 - \sqrt{2}$, $x^3 + \dfrac{5}{6}$ etc.

The family of all antiderivatives of a function $f(x)$ is called the **indefinite integral** of the function $f(x)$ and is denoted by the symbol $\int f(x)\, dx$ whose meaning will be discussed at length in Sec. XIV.2.

Here $\int$ is the **integral sign**, $f(x)$ is the **integrand** and $f(x)\, dx$ is the **element of integration**. Thus,

$$\text{if}\quad F'(x) = f(x)\quad \text{then}\quad \int f(x)\, dx = F(x) + C \qquad \text{and}$$
$$\text{vice versa} \tag{1}$$

For instance, $\int 3x^2\, dx = x^3 + C$. In other words, the indefinite integral is the general expression of antiderivatives which involves an *arbitrary constant*, and every concrete numerical value of the constant yields a certain concrete antiderivative.

Formula (1) implies that

$$\left(\int f(x)\, dx \right)' = f(x), \quad d\left(\int f(x)\, dx \right) = f(x)\, dx,$$
$$\int (dF(x)) = F(x) + C \tag{2}$$

Therefore, the signs of integration and differentiation mutually cancel out. The result of computing an indefinite integral can always be verified by finding the derivative of the result. If the answer is correct the differentiation must yield the integrand. To each formula of differential calculus (see Secs. IV.4-5) there corresponds a certain formula of integral calculus.

2. The Simplest Integrals. These integrals are obtained by reversing formulas for differentiation of basic elementary functions (see Sec. IV.5). For instance, formula $(\sin x)' = \cos x$ implies

$$\int \cos x\, dx = \sin x + C \tag{3}$$

[see formula (1)]. The formula $(\cos x)' = -\sin x$, or $(-\cos x)' = \sin x$, implies

$$\int \sin x\, dx = -\cos x + C$$

Similarly,

$$\int \frac{1}{\cos^2 x}\, dx = \tan x + C \quad \left(\text{this can be written as } \int \frac{dx}{\cos^2 x} = \tan x + C\right),$$

$$\int \frac{dx}{\sin^2 x} = -\cot x + C \quad \text{and} \quad \int \frac{dx}{\sqrt{1-x^2}} = \arcsin x + C \qquad (4)$$

From formula $(\arccos x)' = -\dfrac{1}{\sqrt{1-x^2}}$ we deduce

$$\int \frac{dx}{\sqrt{1-x^2}} = -\arccos x + C \qquad (5)$$

At first glance one can think that the latter formula contradicts the former. But this is not so because formula $\arcsin x + \arccos x = \dfrac{\pi}{2}$ (see Sec. IV.18) and formulas (4) imply that

$$\int \frac{dx}{\sqrt{1-x^2}} = \arcsin x + C = -\arccos x + \frac{\pi}{2} + C = -\arccos x + C_1$$

where $C_1 = \dfrac{\pi}{2} + C$. Thus, the matter is that the right-hand sides of formulas (4) and (5) contain different arbitrary constants. Such a discrepancy in different forms of an answer can appear in many other examples of indefinite integrals. Naturally, in concrete computations one must choose either formula (4) or formula (5); for instance, we can choose formula (4).

Other formulas of differentiation imply

$$\int \frac{dx}{1+x^2} = \arctan x + C \quad \text{and} \quad \int \frac{dx}{x} = \ln x + C$$

A disadvantage of the last formula is that the function $\dfrac{1}{x}$ whose antiderivative is being computed exists both for $x > 0$ and for $x < 0$ whereas the right-hand side is defined only for $x > 0$. But we can easily verify that there is a more general differentiation formula of the form $(\ln |x|)' = \dfrac{1}{x}$. Indeed, we have $|x| = x$ for $x > 0$ and therefore our formula yields the ordinary derivative of the logarithmic function in this case, and we have $|x| = -x$ for $x < 0$ and hence, for this case, we have $(\ln |x|)' = (\ln(-x))' = \dfrac{1}{-x}(-1) = \dfrac{1}{x}$. Therefore, formula

$$\int \frac{dx}{x} = \ln |x| + C \qquad (6)$$

is valid both for $x > 0$ and for $x < 0$.

Further, we deduce

$$\int a^x\,dx = \frac{a^x}{\ln a} + C \quad \left(\text{in particular,} \quad \int e^x\,dx = e^x + C\right)$$

and

$$\int x^{n-1}\,dx = \frac{x^n}{n} + C, \quad \text{that is} \quad \int x^m\,dx = \frac{x^{m+1}}{m+1} + C$$

Of course, the last formula does not hold for $m = -1$ because the denominator vanishes in this case. But for $m = -1$ the integral turns into $\int \frac{dx}{x}$ which is computed by formula (6).

Further, we have

$$\int \cosh x\,dx = \sinh x + C, \quad \int \sinh x\,dx = \cosh x + C,$$

$$\int \frac{dx}{\cosh^2 x} = \tanh x + C \quad \text{and}$$

$$\int \frac{dx}{\sqrt{x^2+1}} = \sinh^{-1} x + C = \ln\left(x + \sqrt{x^2+1}\right) + C$$

(the formula for $\sinh^{-1} x$ was deduced in Sec. I.28).

The formula deduced above can be obtained without hyperbolic functions if we directly put it down in the form

$$\int \frac{dx}{\sqrt{x^2+1}} = \ln\left(x + \sqrt{x^2+1}\right) + C$$

and then differentiate the answer. Moreover, we have $(\ln |u|)' = \frac{1}{u} u'_x$ and thus

$$\int \frac{dx}{\sqrt{x^2+a}} = \ln\left|x + \sqrt{x^2+a}\right| + C \quad (a = \text{const})$$

because

$$\left(\ln\left|x + \sqrt{x^2+a}\right|\right)' = \frac{1}{x + \sqrt{x^2+a}}\left(1 + \frac{2x}{2\sqrt{x^2+a}}\right) = \frac{1}{\sqrt{x^2+a}}$$

where a is a constant of an arbitrary sign.

The above formulas form the table of integrals of the simplest functions (tabular integrals). The reader should put down the table and learn it by heart because it is widely used for computing integrals. In particular, the formulas enable us to find the integrals of certain functions which can be readily reduced to tabular integrals. This is the so-called *direct integration* which is based on these formulas. To find an integral of this type we take a suitable tabular integral and try to change the answer in such a manner that its derivative should be equal to the integrand. Essentially, this method reduces to applying formula (1). For example, in order to find the

integral

$$\int \cos 3x\, dx \qquad (7)$$

it is natural to take formula (3). But the answer $\sin 3x + C$ does not apply since the derivative of $\sin 3x + C$ is equal to $3\cos 3x$ but not to $\cos 3x$ which is required. But if we divide $\sin 3x$ by 3 the derivative will be multiplied by $\frac{1}{3}$. Hence, $\left(\frac{1}{3}\sin 3x + C\right)' =$

$= \cos 3x$, i.e. $\int \cos 3x\, dx = \frac{1}{3}\sin 3x + C$.

In like manner we find

$$\int \frac{dx}{\sqrt{1-(2x+5)^2}} = \frac{1}{2}\arcsin(2x+5)+C, \quad \int \frac{dx}{x-3} =$$

$$= \ln|x-3|+C, \quad \int \frac{dx}{1+\frac{x^2}{2}} = \int \frac{dx}{1+\left(\frac{x}{\sqrt{2}}\right)^2} =$$

$$= \sqrt{2}\arctan\frac{x}{\sqrt{2}}+C \qquad (8)$$

(check up the answers by means of differentiation!).

Generally, if an integral $\int f(x)\, dx = F(x) + C$ has been found then

$$\int f(ax+b)\, dx = \frac{1}{a} F(ax+b)+C$$

where a and b are arbitrary constant numbers.

3. The Simplest Properties of an Indefinite Integral. These properties are implied by the analogous properties of a derivative (see Sec. IV.4). For instance,

$$\int [f(x) \pm \varphi(x)]\, dx = \int f(x)\, dx \pm \int \varphi(x)\, dx \qquad (9)$$

that is the indefinite integral of an algebraic sum is equal to the sum of the integrals of the summands. To prove the property we take the derivatives of the left-hand side and right-hand side and then verify that the results will be equal on the basis of the first formula (2) and on the basis of the well-known property of derivatives which asserts that the derivative of a sum is equal to the sum of the derivatives of the summands. The derivatives being equal, the corresponding functions can differ only in a constant summand. Thus the proof is completed because the constant must not be put down in formula (9) since the integral signs include arbitrary constant summands.

We similarly verify that

$$\int Af(x)\, dx = A \int f(x)\, dx \quad (A = \text{const}) \qquad (10)$$

Thus, a constant factor can be taken outside the integral sign.

A given integral can often be represented in the form of a sum of tabular integrals with the help of formulas (9) and (10). Then we perform termwise integration and thus obtain the answer (this is the decomposition method). Let us consider several examples; we have

$$\int (3x^3 - 2x + 5)\, dx = \int (3x^3)\, dx - \int (2x)\, dx + \int 5dx =$$

$$= 3 \int x^3\, dx - 2 \int x\, dx + 5 \int dx = 3\,\frac{x^4}{4} - 2\,\frac{x^2}{2} + 5x + C =$$

$$= \frac{3}{4}\,x^4 - x^2 + 5x + C \tag{11}$$

(of course, we have put down only one arbitrary constant here because a sum of arbitrary constants is an arbitrary constant),

$$\int \frac{dx}{a^2 + x^2} = \frac{1}{a^2} \int \frac{dx}{1 + \frac{x^2}{a^2}} = \frac{1}{a^2} \int \frac{dx}{1 + \left(\frac{x}{a}\right)^2} =$$

$$= \frac{1}{a^2}\, a \arctan \frac{x}{a} + C = \frac{1}{a} \arctan \frac{x}{a} + C$$

[compare with example (8)] and, similarly,

$$\int \frac{dx}{\sqrt{a^2 - x^2}} = \arcsin \frac{x}{a} + C \quad (a > 0) \tag{12}$$

(verify the answer!).

Other examples are

$$\int \tan^2 x\, dx = \int \frac{\sin^2 x}{\cos^2 x}\, dx = \int \frac{1 - \cos^2 x}{\cos^2 x}\, dx =$$

$$= \int \left(\frac{1}{\cos^2 x} - 1\right) dx = \int \frac{dx}{\cos^2 x} - \int dx = \tan x - x + C,$$

$$\int \frac{1}{x\,(x-1)}\, dx = \int \frac{x - (x-1)}{x\,(x-1)}\, dx = \int \left(\frac{1}{x-1} - \frac{1}{x}\right) dx =$$

$$= \ln|x-1| - \ln|x| + C = \ln \left|\frac{x-1}{x}\right| + C,$$

$$\int \frac{1}{x^2 - a^2}\, dx = \int \frac{(x+a) - (x-a)}{2a\,(x-a)\,(x+a)}\, dx =$$

$$= \frac{1}{2a} \int \left(\frac{1}{x-a} - \frac{1}{x+a}\right) dx = \frac{1}{2a} \ln \left|\frac{x-a}{x+a}\right| + C$$

In computing the last two integrals we have applied a general method of representing a given fraction in the form of a sum of simpler fractions. When using the method we factor the denominator and then try to represent the numerator as a combination of factors entering into the denominator. If this is possible we get a decomposition of the fraction into a sum of several fractions and perform cancellation in each of the summands.

Let us take one more useful example. Let it be necessary to find the integral $\int \sin 5x \cos 3x\, dx$. From trigonometry we know the formula

$$\sin \alpha \cos \beta = \frac{1}{2}\,[\sin (\alpha + \beta) + \sin (\alpha - \beta)]$$

Therefore,

$$\int \sin 5x \cos 3x\, dx = \int \frac{\sin 8x + \sin 2x}{2}\, dx = -\frac{1}{16} \cos 8x - \frac{1}{4} \cos 2x + C$$

In similar circumstances we also utilize the formulas

$$\sin \alpha \sin \beta = \frac{1}{2}\,[\cos (\alpha - \beta) - \cos (\alpha + \beta)],$$

$$\cos \alpha \cos \beta = \frac{1}{2}\,[\cos (\alpha + \beta) + \cos (\alpha - \beta)],$$

$$\sin^2 \alpha = \frac{1 - \cos 2\alpha}{2} \quad \text{and} \quad \cos^2 \alpha = \frac{1 + \cos 2\alpha}{2}$$

For instance,

$$\int \sin^2 3x\, dx = \int \frac{1 - \cos 6x}{2}\, dx = \frac{1}{2}\, x - \frac{1}{12} \sin 6x + C$$

We now mention one more interesting technique based on applying complex functions of a real argument (see Sec. VIII.6) for which all the integration formulas remain valid. Obviously, if we integrate such a function its real part and imaginary part will also be integrated, that is if $f(x) = u(x) + iv(x)$ then

$$\int f(x)\, dx = \int (u(x) + iv(x))\, dx = \int u(x)\, dx + \int iv(x)\, dx =$$

$$= \int (\operatorname{Re} f(x))\, dx + i \int (\operatorname{Im} f(x))\, dx$$

Therefore $\operatorname{Re}\left(\int f(x)\, dx\right) = \int (\operatorname{Re} f(x))\, dx$ and $\operatorname{Im}\left(\int f(x)\, dx\right) = \int (\operatorname{Im} f(x))\, dx$. For instance, this enables us to find the real integral

$$\int e^{ax} \cos bx\, dx = \int \operatorname{Re} [e^{ax} e^{ibx}]\, dx = \operatorname{Re} \int e^{(a+ib)x}\, dx =$$

$$= \operatorname{Re} \frac{e^{(a+ib)x}}{a + ib} + C = \operatorname{Re} \frac{e^{ax} (\cos bx + i \sin bx)(a - ib)}{a^2 + b^2} +$$

$$+ C = e^{ax} \frac{a \cos bx + b \sin bx}{a^2 + b^2} + C$$

by means of Euler's formula (see Sec. VIII.4).

4. Integration by Parts. Unfortunately, there is no formula expressing the integral of a product of functions in terms of the integrals

of the factors. As we know, the derivative of an elementary function is always an elementary function. But the integral of an elementary function may not be an elementary function, and this fact is connected with the above property. (The notion of an elementary function was introduced in Sec. I.18.) For instance, we have the tabular integrals $\int \sin x \, dx$ and $\int \frac{1}{x} \, dx$ but the integral $\int \frac{\sin x}{x} \, dx$ is not expressible in terms of elementary functions. Integrals of this type will be discussed in Sec. 11.

But if we integrate both sides of formula $(uv)' = u'v + uv'$ (see Sec. IV.4) we obtain

$$uv = \int u'v \, dx + \int uv' \, dx$$

that is

$$\int uv' \, dx = uv - \int u'v \, dx \tag{13}$$

or, which is the same,

$$\int u \, dv = uv - \int v \, du \tag{14}$$

Formula (13) or the equivalent formula (14) is called the **formula of integration by parts**. When applying formula (13) we factor the integrand into two factors, i.e. u and v', and then differentiate the first factor and integrate the second. Hence, we pass to an integral in which u' substitutes for u and v for v'. After such a transformation we may arrive at a tabular integral or at an integral which is simpler than the original one.

Take some examples. In calculating the integral $\int x^2 \ln x \, dx$ we see that it is advisable to differentiate $\ln x$ because this yields a power function which is simpler than the logarithmic one. Of course, we must simultaneously integrate the other factor (x^2) but this yields a power function again. Hence, putting $u = \ln x$ and $dv = x^2 \, dx$ we find $u' = \frac{1}{x}$, $v = \frac{x^3}{3}$, and thus we have

$$\int x^2 \ln x \, dx = \ln x \frac{x^3}{3} - \int \frac{x^3}{3} \, d(\ln x) = \frac{x^3}{3} \ln x -$$
$$- \int \frac{x^2}{3} \, dx = \frac{x^3}{3} \ln x - \frac{x^3}{9} + C$$

It should be noted that while computing v we did not put down the corresponding arbitrary constant, that is we did not write $v = \frac{x^3}{3} + C$, because for our aims it was sufficient to obtain a single function v.

Similarly. we often try to differentiate arc tan x and arc sin x because this yields simpler functions.

When computing the integral $\int x^2 \sin 3x\,dx$ we should differentiate the power function since this reduces the exponent by unity. Therefore, integrating by parts twice in this manner we arrive at a tabular integral. At the same time, the differentiation or integration of the sine yields trigonometric functions and thus this factor is neither simplified nor complicated. Hence, we have

$$\int x^2 \sin 3x\,dx = -\frac{x^2}{3}\cos 3x + \int \frac{1}{3}\cos 3x \cdot 2x\,dx$$

(we have used the expressions $u = x^2$ and $dv = \sin 3x\,dx$ here, i.e. $du = 2x\,dx$ and $v = -\frac{1}{3}\cos 3x$). Further, denoting $u = x$ and $dv = \cos 3x\,dx$, i.e. $du = dx$ and $v = \frac{1}{3}\sin 3x$, we finally obtain

$$\int x^2 \sin 3x\,dx = -\frac{x^2}{3}\cos 3x + \frac{2}{3}\left(\frac{x}{3}\sin 3x - \int \frac{1}{3}\sin 3x\,dx\right) =$$

$$= -\frac{1}{3}x^2\cos 3x + \frac{2}{9}x\sin 3x + \frac{2}{27}\cos 3x + C$$

It can happen that after integration by parts we obtain the original integral on the right-hand side but with another coefficient. Then, combining similar terms we can compute the integral. For instance, we compute the integral $\int \sqrt{1-x^2}\,dx$ by introducing $u = \sqrt{1-x^2}$ and $dv = dx$ $\left(\text{i.e. } du = -\frac{x}{\sqrt{1-x^2}}\,dx \text{ and } v = x\right)$:

$$\int \sqrt{1-x^2}\,dx = x\sqrt{1-x^2} + \int \frac{x^2}{\sqrt{1-x^2}}\,dx = x\sqrt{1-x^2} +$$

$$+ \int \frac{(x^2-1)+1}{\sqrt{1-x^2}}\,dx = x\sqrt{1-x^2} - \int \sqrt{1-x^2}\,dx +$$

$$+ \int \frac{1}{\sqrt{1-x^2}}\,dx = x\sqrt{1-x^2} - \int \sqrt{1-x^2}\,dx + \text{arc sin } x$$

Now, transposing the integral thus obtained to the left-hand side we receive

$$2\int \sqrt{1-x^2}\,dx = x\sqrt{1-x^2} + \text{arc sin } x + C$$

where C is a constant (because the expression $\int \sqrt{1-x^2}\,dx$, an indefinite integral, which we have transposed is defined to within a constant addend). Consequently,

$$\int \sqrt{1-x^2}\,dx = \frac{1}{2}x\sqrt{1-x^2} + \frac{1}{2}\text{arc sin } x + C_1$$

where $C_1 = \frac{C}{2}$ is an arbitrary constant.

5. Integration by Change of Variable (by Substitution). Here we are going to describe one of the most widely spread methods of integral calculus based on the formula of differentiation of a composite function (Sec. IV.4). Suppose a function $F(x)$ is an antiderivative of $f(x)$, and let x depend in a certain manner on t, i.e. $x = \varphi(t)$. Compute the derivative of $F(x)$ with respect to t:

$$[F(x)]'_t = [F(x)]'_x \cdot x'_t = f(x)\,\varphi'(t) = f[\varphi(t)]\,\varphi'(t)$$

If we integrate both sides with respect to t we get

$$F(x) + C = \int f[\varphi(t)]\,\varphi'(t)\,dt$$

Therefore, by formula (1), we have

$$\left[\int f(x)\,dx\right]\Big|_{x=\varphi(t)} = \int f[\varphi(t)]\,\varphi'(t)\,dt \qquad (15)$$

It is this formula that is the basic **formula of integration by substitution (by change of variable).**

We have $\varphi'(t)\,dt = dx$, and the right-hand side of formula (15) can therefore be rewritten as $\int f(x)\,dx$. But in the process of integration we do not regard x as an independent variable but consider it to be dependent on t.

Consequently, formula (15) can be interpreted as follows: any integration formula of the form

$$\int f(x)\,dx = F(x) + C \qquad (16)$$

remains valid if we make an arbitrary substitution $x = \varphi(t)$ both in the right-hand side and in the element of integration. Any formula of the form (16) is invariant in this sense.

For instance, substituting $x = u^3$ into formula (3) we obtain

$$\int \cos u^3\,d(u^3) = \sin u^3 + C, \quad \text{that is} \quad \int u^2 \cos u^3\,du = \frac{1}{3}\sin u^3 + C$$

and the like. Of course, when we apply formula (15) to a practical problem we do not start from a tabular formula; on the contrary, we try to find a change of variable which reduces a given integral to a certain tabular integral.

Let us consider some examples. To get rid of the radical in the integral $\int \dfrac{\sqrt{x}}{1+x}\,dx$ we perform the change of variable $x = t^2$ and $dx = 2t\,dt$:

$$\int \frac{\sqrt{x}}{1+x}\,dx = \int \frac{t}{1+t^2}\cdot 2t\,dt = 2\int \frac{t^2+1-1}{1+t^2}\,dt =$$

$$= 2\left(\int dt - \int \frac{dt}{1+t^2}\right) = 2\,(t - \text{arc tan } t + C) = 2\,(\sqrt{x} - \text{arc tan }\sqrt{x}) + C$$

Hence, after performing the substitution and integration we must make the reverse substitution, that is we must pass from t to x.

We sometimes regard the right-hand side of formula (15) as given and apply the formula for computing the left-hand side, that is we make the substitution $\psi(x) = u$ instead of $x = \varphi(t)$. For example, to find the integral $\int xe^{x^2}\,dx$ we take advantage of the fact that the element of integration can be simply expressed in terms of x^2 because $x\,dx = \frac{1}{2}\,d(x^2)$. Therefore, putting $x^2 = u$, $2x\,dx = du$ we obtain

$$\int xe^{x^2}\,dx = \int e^u\,\frac{1}{2}\,du = \frac{1}{2}\,e^u + C = \frac{1}{2}\,e^{x^2} + C$$

Substitutions of the form $\psi(x) = \varphi(t)$ are also applied in some problems.

We could compute integral (7) by means of the substitution $3x = t$, $3dx = dt$:

$$\int \cos 3x\,dx = \int \cos t\,\frac{dt}{3} = \frac{1}{3}\int \cos t\,dt = \frac{1}{3}\sin t + C = \frac{1}{3}\sin 3x + C$$

We can sometimes perform calculations of this type without putting down the change of variable explicitly; for instance,

$$\int \cos 3x\,dx = \int \cos 3x\,\frac{d(3x)}{3} = \frac{1}{3}\int \cos 3x\,d(3x) = \frac{1}{3}\sin 3x + C$$

Here we have used the invariance of formula (3).

Similarly,

$$\int \tan x\,dx = \int \frac{\sin x}{\cos x}\,dx = -\int \frac{d(\cos x)}{\cos x} = -\ln|\cos x| + C$$

In general, we have

$$\int \frac{f'(x)}{f(x)}\,dx = \int \frac{df(x)}{f(x)} = \ln|f(x)| + C \qquad (17)$$

Further, we have

$$\int \frac{x}{\sqrt{x^2+1}}\,dx = \int (x^2+1)^{-\frac{1}{2}} \cdot \frac{1}{2}\,d(x^2+1) =$$

$$= \frac{1}{2} \cdot \frac{(x^2+1)^{\frac{1}{2}}}{\frac{1}{2}} + C = \sqrt{x^2+1} + C$$

Such an integration formula can be put down in the general form as

$$\int \frac{f'(x)}{\sqrt{f(x)}}\,dx = \int [f(x)]^{-\frac{1}{2}}\,df(x) = \frac{[f(x)]^{\frac{1}{2}}}{\frac{1}{2}} + C = 2\sqrt{f(x)} + C \quad (18)$$

In particular, by means of formulas (17) and (18) and by the method of completing a square, we can compute the integrals of the form

$$\int \frac{ax+b}{\sqrt{px^2+qx+r}}\, dx \quad \text{and} \quad \int \frac{ax+b}{px^2+qx+r}\, dx$$

which are widely encountered.

For example, let us illustrate the computation of the integral $\int \frac{2x-3}{\sqrt{-3x^2+2x+1}}\, dx$. To do this we take into account that the derivative of the radicand is equal to $-6x+2 = -6\left(x-\frac{1}{3}\right)$:

$$\int \frac{2x-3}{\sqrt{-3x^2+2x+1}}\, dx = \int \frac{2\left[\left(x-\frac{1}{3}\right)+\frac{1}{3}\right]-3}{\sqrt{-3x^2+2x+1}}\, dx =$$

$$= \int \frac{2\left(x-\frac{1}{3}\right)}{\sqrt{-3x^2+2x+1}}\, dx + \int \frac{-\frac{7}{3}}{\sqrt{-3x^2+2x+1}}\, dx =$$

$$= -\frac{1}{3}\int \frac{-6\left(x-\frac{1}{3}\right)}{\sqrt{-3x^2+2x+1}}\, dx - \frac{7}{3\sqrt{3}}\int \frac{dx}{\sqrt{-\left(x^2-\frac{2}{3}x-\frac{1}{3}\right)}} =$$

$$= -\frac{2}{3}\sqrt{-3x^2+2x+1} - \frac{7}{3\sqrt{3}}\int \frac{d\left(x-\frac{1}{3}\right)}{\sqrt{\frac{4}{9}-\left(x-\frac{1}{3}\right)^2}} =$$

$$= -\frac{2}{3}\sqrt{-3x^2+2x+1} - \frac{7}{3\sqrt{3}}\arcsin \frac{x-\frac{1}{3}}{\frac{2}{3}} + C =$$

$$= -\frac{2}{3}\sqrt{-3x^2+2x+1} - \frac{7}{3\sqrt{3}}\arcsin \frac{3x-1}{2} + C \qquad (19)$$

[see formula (12)].

The problem of integration is much more complicated than the problem of differentiation. The reader should exercise much in order to learn elementary methods of integration.

§ 2. Standard Methods of Integration

Here we shall present some classes of functions which can be integrated by means of certain standard methods. It should be noted that in some cases these standard methods may not be the simplest. It is often advisable to perform certain preliminary trans-

formations or to apply directly methods of § 1 in order to simplify calculations. But the reader will be able to find the simplest technique leading to the desired result only after necessary experience in computing integrals will be acquired.

6. Integration of Rational Functions. Rational functions are integrated on the basis of the results of Sec. VIII.10. As was shown, any rational function (rational fraction) can be represented in the form of a sum of an entire rational function (a polynomial), in case the fraction in question is improper, and partial rational fractions.

A polynomial can be integrated termwise by means of the simplest methods [for instance, see example (11)]. Partial fractions of the form $\dfrac{A}{(x-a)^\alpha}$ can also be integrated quite easily.

For instance, if it is necessary to integrate function (VIII.38) then, by (VIII.39) and (VIII.42), we obtain

$$\int \frac{x^3 - 2x + 3}{x(x-1)(x+2)^2}\,dx = \int \left[-\frac{3}{4}\frac{1}{x} + \frac{2}{9}\frac{1}{(x-1)} - \frac{1}{6}\frac{1}{(x+2)^2} + \right.$$

$$\left. + \frac{55}{36}\frac{1}{(x+2)} \right] dx = -\frac{3}{4}\ln|x| + \frac{2}{9}\ln|x-1| + \frac{1}{6}\frac{1}{(x+2)} +$$

$$+ \frac{55}{36}\ln|x+2| + \text{const}$$

Hence, now we must consider partial fractions of the form

$$\frac{Mx + N}{(x^2 + px + q)^\beta} \qquad (p^2 - 4q < 0) \tag{20}$$

We begin the integration with a simplification of the numerator. Namely, taking into account that $(x^2 + px + q)' = 2\left(x + \dfrac{p}{2}\right)$ we replace x in the numerator by $\left(x + \dfrac{p}{2}\right) - \dfrac{p}{2}$ and then combine similar terms without removing the parentheses. After that we break up the integral into two integrals [as in calculating expression (19)]. The first integral is of the form

$$\int \frac{\left(x + \dfrac{p}{2}\right) dx}{(x^2 + px + q)^\beta} = \frac{1}{2}\int \frac{d(x^2 + px + q)}{(x^2 + px + q)^\beta}$$

and we therefore find it immediately. The second integral is of the form $\int \dfrac{dx}{(x^2 + px + q)^\beta}$. To compute it we complete the square in the denominator which results in $x^2 + px + q = (x + a)^2 + b$ where a and b are constants. Now, if we put $x + a = y$ we arrive at the integral

$$I_\beta = \int \frac{1}{(y^2 + b)^\beta}\,dy \tag{21}$$

which can be easily found in the case $\beta = 1$ (how can we do it?).
To find integral (21) for $\beta = 2, 3, \ldots$ we shall deduce a recurrence
formula which will enable us to pass from I_β to the simpler integral
$I_{\beta-1}$ etc. The formula is obtained by means of integration by parts.
We have

$$I_\beta = \frac{1}{b} \int \frac{b}{(y^2+b)^\beta}\, dy = \frac{1}{b} \int \frac{(b+y^2)-y^2}{(y^2+b)^\beta}\, dy =$$

$$= \frac{1}{b} I_{\beta-1} - \frac{1}{b} \int y \cdot \frac{y}{(y^2+b)^\beta}\, dy$$

Here we put $u = y$, $dv = \dfrac{y}{(y^2+b)^\beta}\, dy$, i.e. $du = dy$, $v = \displaystyle\int \frac{y\, dy}{(y^2+b)^\beta} =$

$$= \frac{1}{2} \int \frac{d\,(y^2+b)}{(y^2+b)^\beta} = \frac{-1}{2\,(\beta-1)} \frac{1}{(y^2+b)^{\beta-1}}\,.$$

Hence,

$$I_\beta = \frac{1}{b} I_{\beta-1} + \frac{y}{2b\,(\beta-1)\,(y^2+b)^{\beta-1}} -$$

$$- \frac{1}{b} \int \frac{1}{2\,(\beta-1)} \frac{1}{(y^2+b)^{\beta-1}}\, dy = \frac{y}{2b\,(\beta-1)\,(y^2+b)^{\beta-1}} + \frac{2\beta-3}{2b\,(\beta-1)} I_{\beta-1} \qquad (22)$$

(let the reader verify all the calculations!).

As we have already mentioned, formulas of this type are called
recurrence formulas. Such formulas express an unknown quantity
dependent on a number (this is the quantity I_β with the number β
in our case) in terms of similar quantities with lower numbers (this
is the quantity $I_{\beta-1}$ with the number $\beta - 1$ in our case). These
formulas may not yield the solution immediately but they enable
us to obtain the solution after several successive reductions of the
number. Thus, formula (22) expresses I_β in terms of $I_{\beta-1}$. If we
repeatedly apply the formula to $I_{\beta-1}$, that is if we substitute $\beta - 1$
for β into formula (22), we obtain the expression of $I_{\beta-1}$ in terms of
$I_{\beta-2}$ etc. Finally, we arrive at the integral I_1 which is immediately
found, as has already been indicated.

It is worth noting that in the above calculations we did not use
the fact that the trinomial in the denominator of expression (20)
has imaginary roots. The procedure can therefore be applied to
integrating a fraction of form (20) when the denominator has real
roots without decomposing the fraction into two summands of the
form $\dfrac{A}{(x-a)^\nu}$.

Integral (21) for $b > 0$ can also be computed with the help of the
substitution $y = \sqrt{b} \tan t$. This leads to an integral of a power
of $\cos t$ which will be discussed later.

Hence, the integral of a rational fraction is always expressible in terms of elementary functions, and this can be achieved by means of the above standard methods. The elementary functions in terms of which an integral of this type is expressed are rational functions, the logarithmic function and the arc tangent. The most difficult thing in the integration is the factorization of the denominator in accordance with formula (VIII.29).

Methods of computing many integrals of other types which we are going to study here are essentially based on the transition from a given integral to an integral of a rational function by means of suitable substitutions. This is the so-called **rationalization** of the integral which reduces the computation to the above standard methods.

7. Integration of Irrational Functions Involving Linear and Linear-Fractional Expressions. First we take an integral of the form

$$\int R\left(x, \sqrt[n]{ax+b}\right) dx \quad (n=2, 3, \ldots) \tag{23}$$

where a and b are constants and $R(x, y)$ is a rational function of its two arguments x and y (see Sec. I.17). The integrand is an irrational function here because it contains the radical. To rationalize the integral let us use the substitution

$$ax + b = t^n, \quad a\, dx = nt^{n-1}\, dt$$

which yields

$$\int R\left(x, \sqrt[n]{ax+b}\right) dx = \int R\left(\frac{t^n-b}{a}, t\right) nt^{n-1}\, dt$$

The integrand in the last integral is a rational function (why?). Similarly, an integral of the form

$$\int R\left(x, \sqrt[n]{ax+b}, \sqrt[m]{ax+b}, \ldots\right) \cdot dx \quad (n, m = 2, 3, 4, \ldots) \tag{24}$$

where $R(x, y, z, \ldots)$ is a rational function of its arguments $x, y, z, \ldots$ goes into an integral of a rational function after the substitution $ax + b = t^p$ with p suitably chosen (how must we choose p in the general case?).

For example, the substitution $2x + 3 = t^6$, $2dx = 6t^5\, dt$ yields

$$\int \frac{dx}{\sqrt{2x+3} - 2\sqrt[3]{2x+3}} = \int \frac{3t^5\, dt}{t^3 - 2t^2} = 3\int \frac{t^3}{t-2}\, dt$$

Performing the division of t^3 by $t-2$ we find

$$\frac{t^3}{t-2} = t^2 + 2t + 4 + \frac{8}{t-2}$$

and hence we finally obtain

$$\int \frac{dx}{\sqrt{2x+3}-2\sqrt[3]{2x+3}} = 3 \int \left(t^2 + 2t + 4 + \frac{8}{t-2}\right) dt =$$

$$= t^3 + 3t^2 + 12t + 24 \ln|t-2| + C = \sqrt{2x+3} + 3\sqrt[3]{2x+3} +$$

$$+ 12\sqrt[6]{2x+3} + 24 \ln\left|\sqrt[6]{2x+3}-2\right| + C \quad \bullet$$

The rationalization of an integral of the form

$$\int R\left(x, \sqrt[n]{\frac{ax+b}{cx+d}}\right) dx \quad (n = 2, 3, \ldots) \tag{25}$$

where $R(x, y)$ is a rational function is carried out by means of the substitution

$$\frac{ax+b}{cx+d} = t^n, \quad ax+b = cxt^n + d \cdot t^n, \quad x = \frac{d \cdot t^n - b}{a - ct^n}$$

Thus, integrals (23)-(25) in which R is a rational function of its arguments are always expressible in terms of elementary functions.

8. Integration of Irrational Expressions Containing Quadratic Trinomials. Here we mean integrals of the form

$$\int R\left(x, \sqrt{ax^2+bx+c}\right) dx \tag{26}$$

where $R(x, y)$ is a rational function of its arguments. Such an integral can also be expressed in terms of elementary functions in all cases. In computing these integrals we apply *trigonometric substitutions*. In order to do this we first complete the square and pass to a new integral:

$$\int R\left(x, \sqrt{ax^2+bx+c}\right) dx = \int R\left(x, \sqrt{\pm(kx+l)^2 \pm m^2}\right) dx$$

where k, l and m are constants. After that we use one of the following substitutions:

$$kx+l = m \tan t \quad \text{for the radical } \sqrt{(kx+l)^2+m^2}$$

$$kx+l = m \sin t \quad \text{for the radical } \sqrt{-(kx+l)^2+m^2} \text{ and}$$

$$kx+l = \frac{m}{\cos t} \quad \text{for the radical } \sqrt{(kx+l)^2-m^2}$$

(of course, we cannot have the case $\sqrt{-(kx+l)^2-m^2}$ for real integrals). The substitutions enable us to extract the roots (check it up!) and thus we come to an integral of the form

$$\int R_1(\cos t, \sin t) dt \tag{27}$$

where $R_1(x, y)$ is another rational function of its arguments. In Sec. 10 we shall describe methods of computing an integral of form (27).

We sometimes use the *hyperbolic substitutions*

$$kx + l = m \sinh t, \quad kx + l = m \tanh t \quad \text{and}$$
$$kx + l = m \cosh t$$

There are certain direct methods that can be applied to computing integrals of form (26). For instance, we can often pass to an integral of the form

$$\int \frac{P_n(x)}{\sqrt{ax^2+bx+c}}\, dx \tag{28}$$

where $P_n(x)$ is a polynomial of the nth degree. Such an integral can be easily found with the help of the method of undetermined coefficients. Let us show that the integral can be represented in the form

$$Q_{n-1}(x)\sqrt{ax^2+bx+c} + K \int \frac{dx}{\sqrt{ax^2+bx+c}} \tag{29}$$

where $Q_{n-1}(x)$ is a polynomial of the $(n-1)$th degree and K is a constant. The last integral is readily found (see the end of Sec. 5).

For definiteness, let $n = 3$. Equating (28) to (29) we obtain

$$\int \frac{\alpha x^3+\beta x^2+\gamma x+\delta}{\sqrt{ax^2+bx+c}}\, dx = (Ax^2+Bx+C)\sqrt{ax^2+bx+c} +$$
$$+ K \int \frac{dx}{\sqrt{ax^2+bx+c}} \tag{30}$$

where all the coefficients on the left-hand side are given and the coefficients A, B, C and K should be determined. In order to find them let us differentiate equality (30):

$$\frac{\alpha x^3+\beta x^2+\gamma x+\delta}{\sqrt{ax^2+bx+c}} = (2Ax+B)\sqrt{ax^2+bx+c} +$$
$$+ (Ax^2+Bx+C)\frac{2ax+b}{2\sqrt{ax^2+bx+c}} + \frac{K}{\sqrt{ax^2+bx+c}},$$
$$\alpha x^3 + \beta x^2 + \gamma x + \delta = (2Ax+B)(ax^2+bx+c) +$$
$$+ \frac{1}{2}(Ax^2+Bx+C)(2ax+b) + K$$

Equating coefficients in equal powers of x, that is coefficients in x^3, x^2, $x^1 = x$, $x^0 = 1$, we find, in succession:

$$\left.\begin{aligned}
3aA &= \alpha \\
\frac{5}{2}bA + 2aB &= \beta \\
2cA + \frac{3}{2}bB + aC &= \gamma \\
cB + \frac{1}{2}bC + K &= \delta
\end{aligned}\right\}$$

We have $a \neq 0$ and therefore A is easily found from the first equation. Substituting the value of A thus found into the second equation we determine B etc. Consequently, we thus can determine all the coefficients A, B, C and K which justifies formula (30).

For instance, let it be necessary to compute the integral

$$I = \int \sqrt{2x^2 - 2x + 1}\, dx$$

In order to do this we write

$$\int \sqrt{2x^2 - 2x + 1}\, dx = \int \frac{2x^2 - 2x + 1}{\sqrt{2x^2 - 2x + 1}}\, dx =$$

$$= (Ax + B)\sqrt{2x^2 - 2x + 1} + K \int \frac{dx}{\sqrt{2x^2 - 2x + 1}},$$

$$\frac{2x^2 - 2x + 1}{\sqrt{2x^2 - 2x + 1}} =$$

$$= A\sqrt{2x^2 - 2x + 1} + (Ax + B)\frac{4x - 2}{2\sqrt{2x^2 - 2x + 1}} + \frac{K}{\sqrt{2x^2 - 2x + 1}}$$

Hence, we have

$$2x^2 - 2x + 1 = A(2x^2 - 2x + 1) + (Ax + B)(2x - 1) + K$$

and therefore

$$\left.\begin{aligned} 4A &= 2 \\ -3A + 2B &= -2 \\ A - B + K &= 1 \end{aligned}\right\} \quad \text{which yields} \quad A = \frac{1}{2}, \quad B = -\frac{1}{4}, \quad K = \frac{1}{4}$$

Thus, we obtain

$$\int \frac{dx}{\sqrt{2x^2 - 2x + 1}} = \frac{1}{\sqrt{2}} \int \frac{dx}{\sqrt{x^2 - x + \frac{1}{2}}} = \frac{1}{\sqrt{2}} \int \frac{d\left(x - \frac{1}{2}\right)}{\sqrt{\left(x - \frac{1}{2}\right)^2 + \frac{1}{4}}} =$$

$$= \frac{1}{\sqrt{2}} \ln\left| x - \frac{1}{2} + \sqrt{\left(x - \frac{1}{2}\right)^2 + \frac{1}{4}} \right| + C =$$

$$= \frac{1}{\sqrt{2}} \ln \frac{\left| 2x - 1 + \sqrt{2x^2 - 2x + 1} \right|}{2} + C =$$

$$= \frac{1}{\sqrt{2}} \ln\left| 2x - 1 + \sqrt{2x^2 - 2x + 1} \right| + C_1$$

where $C_1 = -\frac{1}{\sqrt{2}} \ln 2 + C$.

Finally,

$$I = \left(\frac{1}{2}x - \frac{1}{4}\right) \sqrt{2x^2 - 2x + 1} + \frac{1}{4\sqrt{2}} \ln\left|2x - 1 + \sqrt{2x^2 - 2x + 1}\right| +$$

$$+ C_2, \quad C_2 = \frac{1}{4}C_1$$

Of course, we can simply write C instead of C_2 in the final answer. The integral

$$\int \frac{dx}{(x-\alpha)^n \sqrt{ax^2 + bx + c}} \quad (n = 1, 2, \ldots) \tag{31}$$

can be reduced to integral (28) by means of the substitution $x - \alpha = \frac{1}{t}$. Hence, after the substitution has been carried out we can apply the method of undetermined coefficients. The method can also be applied to an integral of the form

$$\int \frac{Q(x)}{P(x)} \frac{1}{\sqrt{ax^2 + bx + c}} \, dx \tag{32}$$

where $P(x)$ and $Q(x)$ are polynomials. Indeed, if we decompose the fraction $\frac{Q(x)}{P(x)}$ into an entire part and a sum of partial rational fractions of type $\frac{A}{(x-a)^\nu}$ [see formula (VIII.37)] then integral (32) breaks into a sum of integrals of forms (28) and (31).

9. **Integrals of Binomial Differentials.** A binomial differential is an expression of the form

$$(ax^n + b)^p \, x^m \, dx$$

which enters into an integral of the form

$$I = \int (ax^n + b)^p x^m \, dx$$

that we are going to consider here. The numbers n, p and m in the expression are rational numbers, that is they are integers or rational fractional numbers. In 1730 Christian Goldbach (1690-1764), a Russian mathematician, indicated three cases when the integral of a binomial differential can be expressed in terms of elementary functions:

1. *The number p is an integer.* If $p > 0$ we should simply remove the brackets and perform termwise integration. If $p < 0$ we must make the substitution $x = t^k$ and choose k in such a way that all the exponents become integers. This being always possible, we thus get an integral of a rational function.

2. *The number* $\frac{m+1}{n}$ *is an integer.* Substitute $ax^n + b = u$. Then we get

$$I = \int u^p \left[\left(\frac{u-b}{a} \right)^{\frac{1}{n}} \right]^m \cdot \frac{1}{n} \left(\frac{u-b}{a} \right)^{\frac{1}{n}-1} \cdot \frac{1}{a} \, du =$$

$$= \frac{1}{na^{\frac{m+1}{n}}} \int (u-b)^{\frac{m+1}{n}-1} u^p \, du$$

The exponent $\frac{m+1}{n}-1$ being an integer, we thus arrive at preceding case 1.

3. *The number* $\frac{m+1}{n}+p$ *is an integer.* Here we use the substitution $ax^n + b = ux^n$ which again yields case 1. We leave the calculations to the reader.

It was only in 1853 that the prominent Russian mathematician P. L. Chebyshev (1821-1894) proved that the integral of a binomial differential cannot be expressed in terms of elementary functions (see Sec. 11) except for the three cases enumerated above.

10. Integration of Functions Rationally Involving Trigonometric Functions. Here we shall deal with an integral of the form

$$\int R (\sin x, \, \cos x) \, dx \tag{33}$$

where $R(u, v)$ is a rational function in u and v. Such an integral is always expressible in terms of elementary functions. To prove this let us make the so-called *universal substitution* $\tan \frac{x}{2} = t$. Then we have

$$\sin x = \frac{2 \tan \frac{x}{2}}{1 + \tan^2 \frac{x}{2}} = \frac{2t}{1+t^2}, \quad \cos x = \frac{1-t^2}{1+t^2}, \quad dx = \frac{2dt}{1+t^2} \tag{34}$$

(verify the calculations!). Hence, integral (33) reduces to the integral

$$\int R \left(\frac{2t}{1+t^2}, \, \frac{1-t^2}{1+t^2} \right) \frac{2}{1+t^2} \, dt$$

where the integrand is a rational function of t. The last integral can be found by means of the method of Sec. 6.

The universal substitution (34) often leads to very complicated expressions containing rational fractions and it is therefore preferable to avoid it in problem-solving practice. In certain particular cases it is better to use some other substitutions which we are going to consider here.

1. Let the integrand in integral (33) be an odd function with respect to $\sin x$, that is let $R\,(-\sin x,\ \cos x) \equiv -R\,(\sin x,\ \cos x)$. Then we can write

$$I = \int R\,(\sin x,\ \cos x)\,dx = \int \frac{R\,(\sin x,\ \cos x)}{\sin x}\,\sin x\,dx =$$

$$= \int R_1\,(\sin x,\ \cos x)\,\sin x\,dx$$

where R_1 is an even function with respect to $\sin x$. R_1 being a rational function, we can easily express it in terms of $\sin^2 x$ and $\cos x$. It follows that

$$I = \int R_2\,(\sin^2 x,\ \cos x)\,\sin x\,dx = -\int R_2\,(1 - \cos^2 x,\ \cos x)\,d\cos x$$

and therefore if we put $\cos x = t$ we arrive at an integral of a rational function.

2. *Similarly, if the integrand in (33) is an odd function with* respect to $\cos x$ then the substitution $\sin x = t$ rationalizes the integral.

For example,

$$\int \frac{\sin^2 x\,dx}{\cos^3 x - 2\sin x \cos x} = \int \frac{\sin^2 x \cos x\,dx}{\cos^2 x\,(\cos^2 x - 2\sin x)}$$

Putting $\sin x = t$, $\cos x\,dx = dt$ we derive

$$\int \frac{\sin^2 x\,dx}{\cos^3 x - 2\sin x \cos x} = \int \frac{t^2\,dt}{(1 - t^2)\,(1 - t^2 - 2t)}$$

The last integral is readily found if we decompose the integrand into partial fractions or if we take advantage of the equality

$$t^2 = \tfrac{1}{4}\,[(1 - t^2) - (1 - t^2 - 2t)]^2$$

3. If the integrand does not change its value when we simultaneously change the signs of $\sin x$ and $\cos x$, that is if

$$R\,(-\sin x,\ -\cos x) \equiv R\,(\sin x,\ \cos x)$$

then we can apply the substitution $\tan x = t$ (or $\cot x = t$). We can easily verify that this yields the rationalization of the integral in the general case but we are not going to do this here because in every concrete example the advisability of the substitution is confirmed by the results of the calculations.

Take an example:

$$\int \frac{dx}{\sin^2 x \cos^4 x} = \int \frac{dx}{\tan^2 x \cos^6 x}$$

Putting $\tan x = t$, $\cos^2 x = \dfrac{1}{1+t^2}$, $\dfrac{dx}{\cos^2 x} = dt$, we complete the integration:

$$\int \frac{dx}{\sin^2 x \cos^4 x} = \int \frac{(1+t^2)^2}{t^2} \, dt = \int \left(t^2 + 2 + \frac{1}{t^2} \right) dt =$$

$$= \frac{t^3}{3} + 2t - \frac{1}{t} + C = \tan^3 x + 2 \tan x - \cot x + C$$

The same integral can be computed if we represent it as

$$\int \frac{(\sin^2 x + \cos^2 x)^2}{\sin^2 x \cos^4 x} \, dx$$

and remove the brackets in the numerator (check it up!).

Let us separately consider integrals of the form

$$\int \sin^m x \cos^n x \, dx \tag{35}$$

where m and n are arbitrary integers of any sign. In case m is odd the integral belongs to case 1 considered above, and thus it can be found by means of the substitution $\cos x = t$. If n is odd the integral belongs to case 2. Finally, if both m and n are even we have case 3. But the calculations can sometimes be simplified. For instance, if $m \geqslant 0$ and $n \geqslant 0$ and if both m and n are even we can apply the formulas

$$\sin^2 x = \frac{1 - \cos 2x}{2}, \quad \sin x \cos x = \frac{1}{2} \sin 2x \quad \text{and} \quad \cos^2 x = \frac{1 + \cos 2x}{2}$$

For example,

$$\int \sin^2 x \cos^6 x \, dx = \int (\sin x \cos x)^2 \cos^4 x \, dx =$$

$$= \frac{1}{16} \int \sin^2 2x \, (1 + \cos 2x)^2 \, dx =$$

$$= \frac{1}{16} \int \sin^2 2x \, dx + \frac{1}{8} \int \sin^2 2x \cos 2x \, dx + \frac{1}{16} \int \sin^2 2x \cos^2 2x \, dx =$$

$$= \frac{1}{32} \int (1 - \cos 4x) \, dx + \frac{1}{16} \int \sin^2 2x \, d(\sin 2x) + \frac{1}{64} \int \sin^2 4x \, dx =$$

$$= \frac{1}{32} \left(x - \frac{\sin 4x}{4} \right) + \frac{1}{16} \frac{\sin^3 2x}{3} + \frac{1}{128} \int (1 - \cos 8x) \, dx =$$

$$= \frac{5}{128} x - \frac{1}{128} \sin 4x + \frac{1}{48} \sin^3 2x - \frac{1}{1024} \sin 8x + C$$

The same result can be obtained if we express the trigonometric functions in terms of exponential functions by using Euler's formulas (VIII.11).

We sometimes perform integration by parts in computing integrals (35) in order to reduce the positive exponents and increase the negative exponents in powers of $\sin x$ and $\cos x$. For example,

we can put $\cos x = u$ and $dv = \dfrac{\cos x}{\sin^3 x}\, dx$ $\left(\text{that is}\right.$
$du = -\sin x\, dx$ and $v = -\dfrac{1}{2\sin^2 x}\left.\right)$ when integrating the function $\dfrac{\cos^2 x}{\sin^3 x}$:

$$\int \frac{\cos^2 x}{\sin^3 x}\, dx = -\frac{\cos x}{2\sin^2 x} - \int \frac{dx}{2\sin x}$$

Now, making the change of variable $\tan \dfrac{x}{2} = t$ we obtain

$$\int \frac{\cos^2 x}{\sin^3 x}\, dx = -\frac{\cos x}{2\sin^2 x} - \frac{1}{2}\int \frac{2dt\,(1+t^2)^{-1}}{2t\,(1+t^2)^{-1}} =$$
$$= -\frac{\cos x}{2\sin^2 x} - \frac{1}{2}\ln|t| + C = -\frac{\cos x}{2\sin^2 x} - \frac{1}{2}\ln\left|\tan \frac{x}{2}\right| + C$$

[here we have utilized formula (34) in transforming the second integral].

11. General Remarks. Since integration is a much more complicated procedure compared to differentiation the reader must carefully study the basic methods of integration. But, on the other hand, it is inexpedient to carry out complicated calculations every time when it is necessary to compute an integral. It is therefore advisable to use reference books in which the most widely encountered integrals are collected in orderly way. In particular, we refer the reader to [7], [19] and [46].

Many important integrals are not elementary functions, that is they cannot be expressed in terms of finite combinations of the simplest elementary functions which are studied in elementary mathematical courses. For instance, the integral

$$\int \sqrt[3]{x^2+1}\, dx = \int (x^2+1)^{\frac{1}{3}}\, dx$$

belongs to the type considered in Sec. 9. But since here we have $n = 2$, $p = \dfrac{1}{3}$ and $m = 0$ the integral cannot be reduced to those three cases we studied in Sec. 9. Similarly, the integrals

$$\left.\begin{array}{l} \displaystyle\int \sin x \cdot x^\alpha\, dx \\[2mm] \displaystyle\int e^{\pm x} \cdot x^\alpha\, dx \\[2mm] \displaystyle\int \cos x \cdot x^\alpha\, dx \end{array}\right\} \quad (\alpha \neq 0,\ 1,\ 2,\ \ldots)$$

are not expressible in terms of elementary functions, and therefore all the integrals that can be reduced to these integrals cannot be expressed in terms of elementary functions either. Examples of such integrals are

$$\int e^{-x^2}\, dx = \frac{1}{2}\int e^{-u} u^{-\frac{1}{2}}\, du$$

$\left(\text{we have applied the substitution } x^2 = u, \ dx = \dfrac{du}{2\sqrt{u}} \text{ here}\right)$,

$$\int \frac{dx}{\ln x} = \int e^u u^{-1}\, du \quad (x = e^u)$$

$$\int \sin x^2\, dx = \frac{1}{2} \int \sin u \cdot u^{-\frac{1}{2}}\, du \quad (x^2 = u)$$

$$\int \cos x^2\, dx = \frac{1}{2} \int \cos u \cdot u^{-\frac{1}{2}}\, du \quad (x^2 = u)$$

and

$$\int \sqrt{\sin x}\, dx = \int u^{\frac{1}{2}} (1 - u^2)^{-\frac{1}{2}}\, du$$

(the last transformation is carried out by means of the change of variable $\sin x = u, \ dx = \dfrac{du}{\sqrt{1 - u^2}}$; this results in the integral on the right-hand side which belongs to the type considered in Sec. 9 for $n = 2$, $p = -\frac{1}{2}$ and $m = \frac{1}{2}$).

There are wide classes of such non-elementary integrals. For instance, as a rule, integrals of the form

$$\int R\left(x, \sqrt{P_n(x)}\right) dx \tag{37}$$

where R, as before, is the sign of a rational function and $P_n(x)$ is a polynomial of degree $n \geqslant 3$ are not expressible in terms of elementary functions.

In the past the fact that certain integrals cannot be reduced to elementary functions was thought of as a catastrophe. But now we can easily overcome such difficulties. First of all, there are extensive tables of many important non-elementary functions in terms of which very many integrals that cannot be reduced to elementary functions are expressed. In Sec. XIV.12 we shall give examples of such non-elementary functions (special functions) to which all the integrals of form (36) can be reduced. Integral (37) for $n = 3$ and $n = 4$ is called an **elliptic integral**. Such an integral can be expressed in terms of the so-called *elliptic functions* which are thoroughly investigated. These methods and formulas can be found in the reference books we mentioned above. We also refer the reader to [23].

Besides, at present the techniques of computing integrals have become so perfect that the investigation of a function represented in the form of an integral is not more difficult than the investigation of a function represented directly, without the integral signs. Therefore now even when we encounter an integral which is expressible in terms of elementary functions but which involves very complicated combinations of them we usually prefer to deal with the integral representation of the function.

CHAPTER XIV

Definite Integral

In solving many important problems we have to sum up an infinite number of infinitesimal summands. This leads to one of the basic concepts of mathematics, namely to that of the definite integral. Essentially, it is this concept to which all the methods of integration presented in Chapter XIII are applied.

§ 1. Definition and Basic Properties

1. Examples Leading to the Concept of Definite Integral. Let us consider a problem which is reverse to the one considered in the end of Sec. IV.1 that led us to the concept of a derivative. Namely, let us regard the law of variation of the instantaneous velocity of a material point $v = v(t)$ as known and calculate the path length covered during a period of time from $t = \alpha$ to $t = \beta$.

Since we do not suppose that the motion is uniform we cannot compute the path as the product of the velocity by the time taken. We shall therefore apply the following procedure. Let us divide the whole time interval into a large number of small subintervals of time which may not be equal to each other:

$$t_0 = \alpha \leqslant t \leqslant t_1, \quad t_1 \leqslant t \leqslant t_2, \quad \ldots, \quad t_{n-1} \leqslant t \leqslant t_n = \beta$$

where $t_1, \ldots, t_{n-1}$ are some intermediate instances of time which are chosen arbitrarily. If these subintervals are sufficiently small we can regard the motion as being uniform during each of the subintervals without making a considerable error. Hence we can put down the following approximate expression of the path:

$$s \approx v_1 \Delta t_1 + v_2 \Delta t_2 + \ldots + v_n \Delta t_n \tag{1}$$

Here v_k $(k = 1, 2, \ldots, n)$ is one of the values of the instantaneous velocity v attained on the kth subinterval of time, i.e. $v_k = v(\tau_k)$ where $t_{k-1} \leqslant \tau_k \leqslant t_k$. and $\Delta t_k = t_k - t_{k-1}$ is the length of the subinterval. [The reader should pay attention to the difference

where each of the points ξ_h is arbitrarily chosen between x_{h-1} and x_h, that is somewhere on the kth subinterval, and $\Delta x_k = x_k - x_{k-1}$. Now let the lengths Δx_h be infinitely decreased; then the limit to which the integral sum tends in this process is called the **definite integral** of the function $f(x)$ taken over the **interval of integration** $\alpha \leqslant x \leqslant \beta$. The definite integral is denoted as

$$\int\limits_\alpha^\beta f(x)\, dx = \lim \sum_{k=1}^n f(\xi_h)\, \Delta x_h \qquad . \qquad (7)$$

Accordingly, in the examples considered in Sec. 1 we obtain, respectively,

$$s = \int\limits_\alpha^\beta v(t)\, dt, \quad V = \int\limits_\alpha^\beta w(t)\, dt, \quad M = \int\limits_\alpha^\beta \rho(s)\, ds \quad \text{and} \quad S = \int\limits_\alpha^\beta f(x)\, dx \quad (8)$$

The last equality implies the geometric meaning of the definite integral in the case when the function $y = f(x)$ (the integrand) is
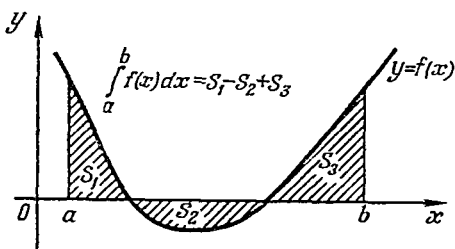
Fig. 245

positive: in this case the integral is equal to the area of the curvilinear trapezoid bounded by the graph of the function, the axis of abscissas and the straight lines parallel to the axis of ordinates passing through the end-points of the interval of integration. The end-points, that is the numbers α and β, are called, respectively, the **lower limit** and the **upper limit of integration.** The expression $f(x)\, dx$ is called the **element of integration.**

If the integrand is negative or changes its sign then some terms entering into integral sum (6) will be negative. Therefore after passing to the limit we see that the integral is equal to the algebraic sum of the areas of the parts of the curvilinear trapezoid which lie over and under the x-axis (see Fig. 245). The areas of the parts lying over the x-axis are taken with the sign $+$ and those under the x-axis are taken with the sign $-$.

Comparing formulas (8) we can also conclude that in order to calculate the path length covered by a point in its rectilinear ·

motion, for a given relationship between the velocity and the time taken which is represented by a graph (see Fig. 246), it is sufficient to compute the area of the corresponding curvilinear trapezoid. In this example we must also take the area with the sign — if $v < 0$, that is if the graph lies under the t-axis, because the increment of the coordinate of the moving point is negative in such a case. This rule of signs for computing an area holds for a great number of other examples.

Let us dwell in more detail on the passage to the limit in formula (7). The limit is sometimes said to be taken as $n \to \infty$ but this is not precise because we do not suppose that the subintervals Δx_k are of the same length and therefore if we limit ourselves to
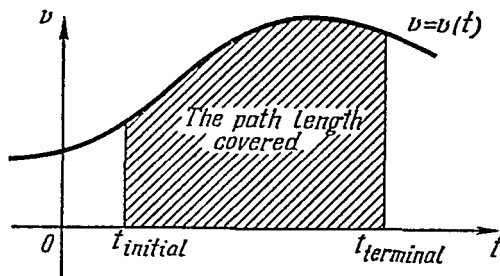
Fig. 246

the condition that only $n \to \infty$ then we can encounter a case when the subintervals belonging to one part of the interval $\alpha \leqslant x \leqslant \beta$ decrease whereas the others do not. It is therefore better to say that the limit is taken while the lengths of the subintervals are infinitely decreased. The degree of the decrease can be characterized by the largest of the lengths Δx_k of a given partition because if the largest length is small then the other lengths are automatically small. Hence, we can say that the passage to the limit in formula (7) is performed in a process in which $\max_k \Delta x_k \to 0$.

Let us consider an example of calculating a definite integral on the basis of its definition (7). Let it be necessary to compute the integral] $\int_0^1 x^2 \, dx$. Divide the interval of integration into five equal parts of length 0.2. For definiteness, let us choose a point on each of the subintervals at its left end-point. Then $\xi_1 = 0.0$, $\xi_2 = 0.2$, $\xi_3 = 0.4$, $\xi_4 = 0.6$, $\xi_5 = 0.8$ and

$$\int_0^1 x^2 \, dx \approx \sum_{k=1}^5 \xi_k^2 \Delta x_k = (0.0^2 + 0.2^2 + 0.4^2 + 0.6^2 + 0.8^2) \cdot 0.2 = 0.24$$

An analogous division into 10 subintervals would yield the value 0.29 and the division into 100 equal subintervals would yield the value 0.33. (In Sec. 3 we shall find the exact value of the integral which turns out to be equal to $\frac{1}{3}$; thus the above value 0.33 is very close to the exact value. We suggest that the reader should elucidate the fact that the approximate values obtained in our example are smaller than the exact value, by means of a graphical construction.)

We see that there is arbitrariness in forming an integral sum because it depends both on the choice of the points of division x_k and on the choice of intermediate points ξ_k. But nevertheless if we take a partition whose subintervals are sufficiently small the sum will be practically equal to its limit, that is to integral (7) which, of course, depends neither on the points x_k nor on the points ξ_k. Each of the summands entering into an integral sum becomes very small when we take a partition in which Δx_k are sufficiently small, the smallness of the summands being implied by the smallness of Δx_k. But at the same time the number of summands becomes so large that the whole sum has a finite value. Roughly speaking, if the number of summands entering into an integral sum is equal to n then each Δx_k (and therefore each of the summands) is of the order of $\frac{1}{n}$ whereas the whole sum is of the order of $n \cdot \frac{1}{n} = 1$, i.e. it is finite. Thus, taking into account that we pass to the limit in formula (7) we can say that *the definite integral is a sum of infinitely many infinitesimal summands.* Practically, we can often regard a definite integral as a sum of a great number of very small homogeneous summands (that is the summands of the same dimension, of the same character, of the same sense etc.), the summands being so small that the sum is practically equal to its limit. Such an approach completely corresponds to the practical concept of infinitely large and infinitely small quantities which are understood (see Secs. III.1 and III.3) as quantities that are, respectively, very large and very small but finite, theoretically. It should be noted that not every sum of infinite number of infinitesimal summands yields an integral. Indeed, as we have seen, for such a sum to assume a finite value, the number and the magnitude of the summands should be coherent in a certain sense.

The interpretation of the integral as a sum accounts for its notation. In fact, if we regard the summands entering into sum (6) as infinitesimals and denote the lengths of the subintervals as $\Delta x_k = dx$ then the whole sum (6) can be put down in the form

$$\sum_{(\text{from } x=\alpha)}^{(\text{to } x=\beta)} f(x)\, dx$$

At first sums were denoted by the letter S. Then it was gradually lengthened and this resulted in modern notation (7).

In conclusion note that an integrand can be either a continuous function on the interval of integration or a discontinuous one, that is having points of discontinuity. But in § 1 we impose the condition that the interval of integration is finite and that the integrand does not approach infinity on the interval. It can be shown that under these assumptions (and under some additional requirements) the definite integral exists, that is it has a finite value. A rigorous proof of this assertion which is not based on physical or geometric considerations can be found in more comprehensive courses on mathematical analysis. As we shall show in § 4, the integral may not have a certain numerical value if the above conditions are violated.

3. **Relationship Between Definite Integral and Indefinite Integral.** We begin with a simple remark that a definite integral does not depend on the notation of the variable of integration, i.e.

$$\int_\alpha^\beta f(x)\,dx = \int_\alpha^\beta f(t)\,dt = \int_\alpha^\beta f(s)\,ds = \ldots \tag{9}$$

Indeed, for example, this is implied by the fact that all the integrals put down above are equal to the same area. Thus, the variable of integration in a definite integral is a *dummy variable* similar to an index of summation (see Sec. III.6) and it can be denoted by any letter or symbol.

Let $f(x)$ be a function that we are going to integrate. But let the lower limit alone (denoted as x_0) be fixed, and the upper limit (which we denote by x) be arbitrary, i.e. variable. Then the value of the integral itself will depend on x, and we can therefore denote it as $\Phi(x)$. Hence we can write

$$\Phi(x) = \int_{x_0}^x f(x)\,dx \quad (x_0 = \text{const})$$

or, taking into account equalities (9),

$$\Phi(x) = \int_{x_0}^x f(t)\,dt \quad (x_0 = \text{const}) \tag{10}$$

The first form of writing may sometimes lead to misunderstandings because the letter x entering into it is simultaneously understood in two different senses, namely as the variable of integration and as the upper limit. Therefore, although the first form is admissible, the second is preferable.

Now let us prove that the function $\Phi(x)$ thus constructed is an antiderivative (see Sec. XIII.1) of the integrand $f(x)$, that is

$$\frac{d}{dx}\int_{x_0}^{x} f(t)\,dt = \left(\int_{x_0}^{x} f(t)\,dt\right)'_x = f(x)$$

Thus, we assert that the derivative of a definite integral with respect to the upper limit is equal to the value of the integrand for
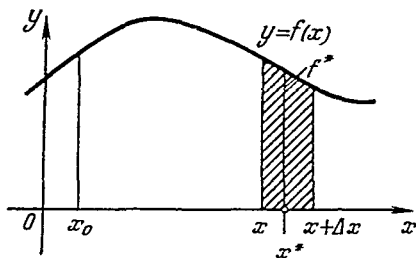
Fig. 247

the value of its argument equal to the upper limit. To prove the assertion we first suppose that $f(x)$ is a continuous function and consider Fig. 247. The geometric meaning of the integral implies that if x gains an increment Δx then $\Delta\Phi$ is equal to the area shaded in Fig. 247. This area is approximately equal to the product $\Delta x \cdot f^*$ where f^* is an intermediate ordinate equal to one of the values of the function taken between the points x and $x + \Delta x$. It follows that $\frac{\Delta\Phi}{\Delta x} = f^* = f(x^*)$. Hence, if $\Delta x \to 0$ then $x^* \to x$, and therefore passing to the limit we obtain
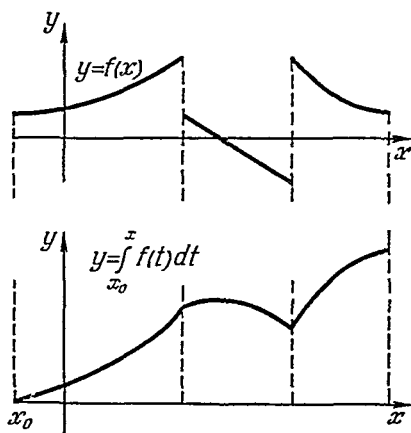
Fig. 248

$$\Phi'(x) = \lim_{\Delta x \to 0} \frac{\Delta\Phi}{\Delta x} = \lim_{\Delta x \to 0} f(x^*) = f(x)$$

which is what we set out to prove.

In particular, we see that *a continuous function always has an antiderivative* (see Sec. XIII.1). To find one of the antiderivatives we can evaluate the definite integral of the given function for a fixed lower limit and regard it as a function of its upper limit.

Now suppose that the integrand is discontinuous (but finite since for the time being we consider only finite functions). Then function (10) is continuous at the points of discontinuity of $f(x)$ but the derivative of $\Phi(x)$ has jump discontinuities at these points. Hence, the graph of $\Phi(x)$ is "broken" at such points (see Fig. 248). We extend the notion of an antiderivative when we admit these "breaks" because at a point of this kind there is no single value of the derivative. But this natural extension enables us to say that each function that is finite everywhere has an antiderivative which is a continuous function.

Now suppose that we have to evaluate the integral

$$I = \int_{\alpha}^{\beta} f(x)\, dx$$

and that we know one of the antiderivatives of the function $f(x)$ which we designate as $F(x)$. The function $\int_{\alpha}^{x} f(t)\, dt$ also being an antiderivative of $f(x)$, we have, by Sec. XIII.1, the relation

$$\int_{\alpha}^{x} f(t)\, dt = F(x) + C$$

where C is a constant. Putting here $x = \alpha$ we see that the geometric meaning of the integral implies that the left-hand side of the relation vanishes, that is we have

$$0 = F(\alpha) + C, \quad C = -F(\alpha) \quad \text{and} \quad \int_{\alpha}^{x} f(t)\, dt = F(x) - F(\alpha)$$

Putting $x = \beta$ in the last formula we obtain, on the basis of relation (9), the formula

$$\int_{\alpha}^{\beta} f(x)\, dx = F(\beta) - F(\alpha) \tag{11}$$

Thus, a definite integral is equal to the increment of an antiderivative of the integrand corresponding to the variation of the independent variable from the lower limit of integration to the upper limit. The right-hand side of formula (11) is also designated as $F(x)\,|_{\alpha}^{\beta}$ where $|_{\alpha}^{\beta}$ is the *sign of double substitution* which means that the lower limit and the upper limit must be substituted for the argument into the function and then the first result subtracted from the second.

Another way of writing formula (11) is

$$\int_\alpha^\beta f(x)\,dx = \left(\int f(x)\,dx\right)\Big|_\alpha^\beta \tag{12}$$

Formula (12) is justified by the fact that

$$\left(\int f(x)\,dx\right)\Big|_\alpha^\beta = (F(x)+C)\Big|_\alpha^\beta = [F(\beta)+C]-[F(\alpha)+C] =$$

$$= F(\beta) - F(\alpha) = \int_\alpha^\beta f(x)\,dx$$

where $F(x)$ is one of the antiderivatives and C is a constant (see Sec. XIII.1).

Thus, a definite integral is equal to the increment of the corresponding indefinite integral. This result is one of the most important theorems in mathematics. It is called the **Newton-Leibniz theorem.**

Let us take an example:

$$\int_0^1 x^2\,dx = \left(\int x^2\,dx\right)\Big|_0^1 = \left(\frac{x^3}{3}\right)\Big|_0^1 = \frac{1^3}{3} - \frac{0^3}{3} = \frac{1}{3}$$

It should be noted that when evaluating the indefinite integral here we have not written the arbitrary constant C because, as it was shown, the terms $+C$ and $-C$ always cancel out.

We see that if the limits of integration are given a definite integral is a constant number whereas the corresponding indefinite integral is a function.

Up to now we assumed that $\alpha < \beta$. Let us extend formula (11) to the case $\alpha \geqslant \beta$. This means that for $\alpha \geqslant \beta$ we regard formula (11) as the definition of the integral written on the left-hand side.

Since $f(x) = F'(x)$ formula (11) can be rewritten as

$$\int_\alpha^\beta F'(x)\,dx = F(\beta) - F(\alpha) \tag{13}$$

Hence, the definite integral of a derivative is equal to the increment of the antiderivative.

4. Basic Properties of Definite Integral.

1. The interchange of the limits of integration yields the multiplication of the integral by -1. Actually, by formula (11), we have

$$\int_\beta^\alpha f(x)\,dx = F(\alpha) - F(\beta) = -[F(\beta) - F(\alpha)] = -\int_\alpha^\beta f(x)\,dx$$

This simple property can also be put down in the form $F(x)|_\alpha^\beta =$ $= -F(x)|_\beta^\alpha$ which enables us to substitute the limits in reverse order if we change the sign of the indefinite integral beforehand. For instance,

$$\int_3^5 \frac{1}{x^2}\,dx = -\frac{1}{x}\Big|_3^5 = \frac{1}{3} - \frac{1}{5} = \frac{2}{15}$$

In particular, property 1 implies the following rule of differentiating an integral with respect to its lower limit:

$$\frac{d}{dx}\left(\int_x^{x_0} f(t)\,dt\right) = \left(\int_x^{x_0} f(t)\,dt\right)'_x = -\frac{d}{dx}\left(\int_{x_0}^x f(t)\,dt\right) = -f(x)$$

2. If the limits of integration coincide then the integral is equal to zero, i.e.

$$\int_\alpha^\alpha f(x)\,dx = 0$$

Property 2 has been already used (see Sec. 3).

3. The theorem on "partition of the interval of integration":

$$\int_\alpha^\beta f(x)\,dx + \int_\beta^\gamma f(x)\,dx = \int_\alpha^\gamma f(x)\,dx$$

for any α, β and γ. In fact, the left-hand side is equal to

$$[F(\beta) - F(\alpha)] + [F(\gamma) - F(\beta)] = F(\gamma) - F(\alpha)$$

which equals $\int_\alpha^\gamma f(x)\,dx.$

4. The integral of a sum of functions is equal to the sum of the integrals of the summands (the same is true for the difference)

$$\int_\alpha^\beta [f(x) \pm \varphi(x)]\,dx = \int_\alpha^\beta f(x)\,dx \pm \int_\alpha^\beta \varphi(x)\,dx$$

To prove the property we apply the analogous property of indefinite integrals (see Sec. XIII.9) and equate the increment of the right-hand side to the increment of the left-hand side as x varies from α to β. The following property is proved in a similar way.

5. A constant factor can be taken outside the sign of the integral:

$$\int_\alpha^\beta Mf(x)\,dx = M\int_\alpha^\beta f(x)\,dx \quad (M = \text{const})$$

Properties 4 and 5 can be formulated simultaneously as "a definite integral is linear with respect to the integrand".

The term "linear" is understood here in the sense of Sec. XI.6. Namely, the formula

$$\int_{\alpha}^{\beta} f(x)\, dx = I \tag{14}$$

for fixed α and β determines a correspondence between the finite (integrable) functions defined over $\alpha \leqslant x \leqslant \beta$ and real numbers I, that is to each function there corresponds a certain number I. In other words, formula (14) defines a mapping of the infinite-dimensional linear space of such functions into the one-dimensional space of all real numbers. Properties 4 and 5 are then nothing but the condition that the mapping is linear. (For instance, let the reader verify that, for $\alpha = 1$ and $\beta = 2$, the number $I = \frac{7}{3}$ corresponds to the function $y = x^2$ and the number $\frac{3}{8}$ corresponds to the function $y = \frac{1}{x^3}$ whereas the number $5 \cdot \frac{7}{3} - 3 \cdot \frac{3}{8} = 10.54$ corresponds to the function $y = 5x^2 - \frac{3}{x^3}$.) A rule, a law, according to which to functions there correspond numbers, is called a **functional**. Hence, formula (14) determines a **linear functional** defined over the above functional space.

6. The formula of integration by parts

$$\int_{\alpha}^{\beta} uv'\, dx = (uv)\Big|_{x=\alpha}^{x=\beta} - \int_{\alpha}^{\beta} u'v\, dx$$

is also deduced from the corresponding formula for indefinite integrals, namely from formula (XIII.13).

Take an instance:

$$\int_{0}^{\pi} x \sin x\, dx = (-x \cos x)\Big|_{0}^{\pi} + \int_{0}^{\pi} \cos x\, dx =$$

$$= (-x \cos x)\Big|_{0}^{\pi} + (\sin x)\Big|_{0}^{\pi} = \pi$$

(here we have put $u = x$, $dv = \sin x\, dx$, $du = dx$ and $v = -\cos x$).

7. The formula of integration by change of variable in definite integrals is obtained if we equate the increments of both sides of formula (XIII.15) corresponding to the variation of t from α to β. Doing this and taking into account that the variable x which equals

$\varphi(t)$ varies from $\varphi(\alpha)$ to $\varphi(\beta)$, we obtain

$$\int_{\alpha}^{\beta} f[\varphi(t)]\,\varphi'(t)\,dt = \left[\int f(x)\,dx\right]\Big|_{x=\varphi(\beta)} - \left[\int f(x)\,dx\right]\Big|_{x=\varphi(\alpha)}$$

Now, taking advantage of formula (12), we finally derive

$$\int_{\alpha}^{\beta} f[\varphi(t)]\,\varphi'(t)\,dt = \int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)\,dx$$

Hence, in applying the formula we should additionally change the limits of integration. To do this we must find an interval which should be run by the new variable so that the old variable of integration should vary over the interval that was originally set for it.

For example, if we want to make the substitution $x = R\sin t$ in order to evaluate the integral

$$\int_{0}^{R} \sqrt{R^2 - x^2}\,dx$$

we must take into account that for x to vary from 0 to R, it is sufficient that t should run from 0 to $\frac{\pi}{2}$. Therefore

$$\int_{0}^{R} \sqrt{R^2 - x^2}\,dx = \int_{0}^{\frac{\pi}{2}} \sqrt{R^2 - R^2\sin^2 t}\,R\cos t\,dt =$$

$$= R^2 \int_{0}^{\frac{\pi}{2}} \cos^2 t\,dt = \frac{R^2}{2} \int_{0}^{\frac{\pi}{2}} (1 + \cos 2t)\,dt =$$

$$= \frac{R^2}{2}\left(t + \frac{\sin 2t}{2}\right)\Big|_{0}^{\frac{\pi}{2}} = \frac{\pi R^2}{4}$$

As we see, in contrast to the change of variable in an indefinite integral, the inverse substitution, that is the transition from the new variable to the old one in the final answer, is not needed in computing a definite integral. We suggest that the reader should construct a geometric figure whose area is expressed by the above integral and, in addition, obtain the same result by means of the substitution $x = R\cos t$.

We have deduced properties 3-5 of the definite integral on the basis of formula (11). But the same could have been achieved with the

help of definition (7) which introduces the definite integral as the limit of the corresponding integral sum. For instance, passing to the limit in the formula

$$\sum_{k=1}^{n} [f(\xi_k) \pm \varphi(\xi_k)] \Delta x_k = \sum_{k=1}^{n} f(\xi_k) \Delta x_k \pm \sum_{k=1}^{n} \varphi(\xi_k) \Delta x_k$$

as the subintervals of the partitions of the interval $\alpha \leqslant x \leqslant \beta$ tend to zero, we receive property 4 etc. Property 8 is implied by the same definition.

8. If the variables in question have certain dimensions then

$$\left[\int_{\alpha}^{\beta} f(x)\, dx \right] = [f] \cdot [x]$$

since the operation of summation and the operation of passing to a limit do not change dimensions.
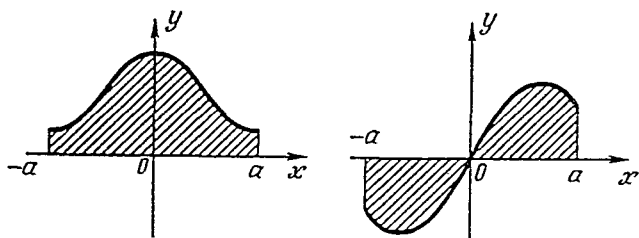
Fig. 249

9. There are certain cases when the integration in symmetric limits can be simplified. Namely, as we see in Fig. 249, we have

$$\int_{-a}^{a} f(x)\, dx = 2 \int_{0}^{a} f(x)\, dx$$

for every even function $f(x)$, and we have

$$\int_{-a}^{a} f(x)\, dx = 0$$

for every odd function $f(x)$.

10. An integral of a periodic function taken over an interval whose length is equal to the period of the function does not depend on the position of the interval on the axis of the variable of integration.

In other words, if $f(x + A) \equiv f(x)$ then the integral

$$I = \int_{x}^{x+A} f(s)\, ds$$

is independent of x. Indeed, according to the rule of differentiating a composite function, and on the basis of the formulas for the derivatives of an integral with respect to its lower and upper limits, we get

$$\frac{dI}{dx} = f(x+A)\frac{d(x+A)}{dx} - f(x)\frac{dx}{dx} = f(x+A) - f(x) \equiv 0$$

(Let the reader prove the property by taking advantage of the geometric meaning of the definite integral.)

In conclusion, let us consider several examples of incorrect evaluation of definite integrals.

$$1.\ \int_{-1}^{1} \sqrt{1-x^2}\, dx = \int_{\pi}^{2\pi} \sqrt{1-\cos^2 t}\,(-\sin t)\, dt =$$

$$= -\int_{\pi}^{2\pi} \sin^2 t\, dt = -\int_{\pi}^{2\pi} \frac{1-\cos 2t}{2}\, dt =$$

$$= -\frac{1}{2}\left(t - \frac{\sin 2t}{2}\right)\Big|_{\pi}^{2\pi} = -\frac{\pi}{2}$$

(we have performed the substitution $x = \cos t$, $\pi \leqslant t \leqslant 2\pi$). The result is apparently incorrect since an integral of a positive function taken in the positive direction (that is from a smaller limit to a larger one) must be positive. The mistake lies in the replacement of $\sqrt{\sin^2 t}$ by $\sin t$ whereas it should have been replaced by $|\sin t|$ (see the end of Sec. I.5). Actually, we have $\sin t < 0$ for $\pi < t < 2\pi$ and hence $\sqrt{\sin^2 t} = -\sin t$ for such t. If we took $\sqrt{\sin^2 t} = |\sin t|$ we should obtain the correct result $\int_{-1}^{1} \sqrt{1-x^2}\, dx = \frac{\pi}{2}$.

We sometimes encounter integrals of the form

$$\int_{a}^{b} |f(x)|\, dx$$

in analogous situations. These integrals can be treated in the following way. We begin with determining the intervals of retention of

the sign of the function $f(x)$ (see Sec. III.15). For instance, let $f(x) > 0$ for $a < x < c$, $f(x) < 0$ for $c < x < d$ and $f(x) > 0$ for $d < x < b$. Then

$$\int_a^b |f(x)|\,dx = \int_a^c |f(x)|\,dx + \int_c^d |f(x)|\,dx + \int_d^b |f(x)|\,dx =$$
$$= \int_a^c f(x)\,dx - \int_c^d f(x)\,dx + \int_d^b f(x)\,dx$$

etc.

2. The correct value of the integral $I = \int_{-1}^{2} x^2\,dx$ is 3 which is obtained without any substitution:

$$\int_{-1}^{2} x^2\,dx = \frac{x^3}{3}\Big|_{-1}^{2} = \frac{2^3}{3} - \frac{(-1)^3}{3} = \frac{8}{3} + \frac{1}{3} = 3$$

But the following calculations are incorrect, and their result contradicts the correct value $I = 3$:

$$\int_{-1}^{2} x^2\,dx = \int_{1}^{4} t\,\frac{1}{2\sqrt{t}}\,dt = \frac{1}{2}\int_{1}^{4}\sqrt{t}\,dt = \frac{1}{2}\frac{t^{\frac{3}{2}}}{\frac{3}{2}}\Big|_{1}^{4} = \frac{8}{3} - \frac{1}{3} = \frac{7}{3}$$

(we have made the change $x^2 = t$, i.e. $x = \sqrt{t}$). The mistake lies in the fact that the formula $x = \sqrt{t}$ of transition from x to t makes no sense for $x < 0$. Therefore. if there are reasons that make it necessary to perform the substitution $x^2 = t$, we must break the integral into two summands according to the formula $\int_{-1}^{2} x^2\,dx =$

$$= \int_{-1}^{0} x^2\,dx + \int_{0}^{2} x^2\,dx$$ and then put $x = -\sqrt{t}$ $(1 \geqslant t \geqslant 0)$ in the first integral and $x = \sqrt{t}$ $(0 \leqslant t \leqslant 4)$ in the second integral (let the reader do it!).

3. As in example 1, the following result is **apparently incorrect**:

$$\int_{-1}^{2} x^{-2}\,dx = \frac{x^{-1}}{-1}\Big|_{-1}^{2} = \frac{2^{-1} - (-1)^{-1}}{-1} = -\frac{3}{2}$$

Indeed, the integrand approaches infinity at $x = 0$, and we cannot therefore apply the Newton-Leibniz formula here. We shall discuss integrals of this type in Sec. 16.

5. Integrating Inequalities. The definition of the integral and its geometric meaning (see Sec. 2) imply that

$$\text{if} \quad f(x) \geqslant 0 \quad \text{and} \quad \alpha < \beta \quad \text{then} \quad \int_\alpha^\beta f(x)\, dx \geqslant 0 \qquad (15)$$

The last inequality turns into the equality if and only if $f(x) \equiv 0$ on the interval $\alpha \leqslant x \leqslant \beta$ in the case of a continuous function $f(x)$. But if we consider discontinuous integrable functions as well then the integral of a function which is different from zero at a finite number of discrete points is nevertheless equal to zero because such points do not affect the value of the integral.

If there is a condition

$$\varphi(x) \leqslant \psi(x) \quad \text{for} \quad \alpha \leqslant x \leqslant \beta \qquad (16)$$

then putting $\psi(x) - \varphi(x) = f(x)$ and taking advantage of assertion (15) we deduce

$$\int_\alpha^\beta [\psi(x) - \varphi(x)]\, dx \geqslant 0 \quad \text{and} \quad \int_\alpha^\beta \psi(x)\, dx - \int_\alpha^\beta \varphi(x)\, dx \geqslant 0$$

Hence, we have

$$\int_\alpha^\beta \varphi(x)\, dx \leqslant \int_\alpha^\beta \psi(x)\, dx \qquad (17)$$

Thus, inequality (16) implies inequality (17) which means that the sign of inequality is retained when we integrate an inequality in the positive direction. (Think about the changes that must be made in the assertion if an inequality is integrated in the negative direction.)

As above, if inequality (16) holds then inequality (17) turns into the equality if and only if $\varphi(x) \equiv \psi(x)$ for $\alpha \leqslant x \leqslant \beta$ in the case of continuous functions $\varphi(x)$ and $\psi(x)$ although in the case of discontinuous integrable functions we can have the equality even if $\varphi(x)$ is different from $\psi(x)$ at separate points.

As a consequence of inequality (17) we obtain a crude estimation of the definite integral: let

$$f_{min} \leqslant f(x) \leqslant f_{max} \qquad (\alpha \leqslant x \leqslant \beta)$$

where f_{min} and f_{max} are two constants. Then integrating |these inequalities we obtain

$$f_{min}(\beta - \alpha) \leqslant \int_\alpha^\beta f(x)\, dx \leqslant f_{max}(\beta - \alpha) \qquad (18)$$

In connection with this estimation let us consider the important notion of the **mean value** of a function which is also called "the *arithmetic mean*" of the function. If a function $f(x)$ is regarded as being defined over an interval $\alpha \leqslant x \leqslant \beta$ then its mean value on the interval is a constant $\bar{f}$ such that the integral of the constant over the interval $\alpha \leqslant x \leqslant \beta$ is equal to the integral of the function taken from α to β. Thus, $\int\limits_{\alpha}^{\beta} \bar{f}\, dx = \int\limits_{\alpha}^{\beta} f(x)\, dx$ i.e.

$$\int\limits_{\alpha}^{\beta} f(x)\, dx = \bar{f} \cdot (\beta - \alpha)$$

The last formula (the first mean value theorem) implies the following expression for the mean value:

$$\bar{f} = \frac{1}{\beta - \alpha} \int\limits_{\alpha}^{\beta} f(x)\, dx \tag{19}$$

As would be expected, inequality (18) implies that

$$f_{min} \leqslant \bar{f} \leqslant f_{max}$$

The geometric meaning of the mean of a function is illustrated in Fig. 250. We see that $\bar{f}$ satisfies the condition that the area of the rectangle $AB'C'D$ is equal to the area of the curvilinear trapezoid $ABCD$. It is clear that if the function $f(x)$ is continuous it takes on the value $\bar{f}$ at a point belonging to the interval $\alpha \leqslant x \leqslant \beta$ (at the point γ in Fig. 250)*. A discontinuous function may not assume its mean value.
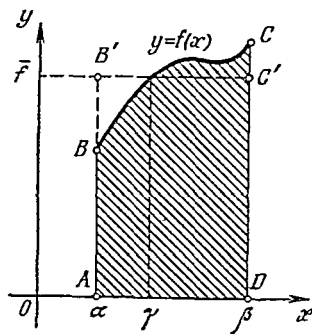
The advisability of the above definition of the mean of a function is well seen if we consider an example of a functional relationship between the instantaneous velocity of a non-uniform motion of a point and a current moment of time. The integral of the velocity over a time interval being equal to the distance passed [see the first formula (8)], we see that the mean value of the velocity on a time

Fig. 250

---

* Hence, for a continuous function $f(x)$, the first mean value theorem is written in the form $\int\limits_{\alpha}^{\beta} f(x)\, dx = f(\gamma)(\beta - \alpha)$ where γ is a certain point belonging to the interval $\alpha \leqslant x \leqslant \beta$. — *Tr.*

interval is a constant velocity such that if the point moved uniformly with this velocity it would cover the same distance as in its non-uniform motion in question, during the same time interval. In other words, formula (19) implies that the mean value of the velocity on a finite time interval is equal to the ratio of the distance covered to the time taken. Hence, this notion agrees with the well-known notion of a mean velocity. The notions of a mean density, mean power etc. are also in agreement with the general notion of the mean value of a function.

If a function is defined over an infinite interval, for instance, in the interval $\alpha \leqslant x < \infty$, then its mean value is defined as

$$\bar{f} = \lim_{\beta \to \infty} \frac{1}{\beta - \gamma} \int_{\gamma}^{\beta} f(x)\, dx \qquad (\gamma = \text{const}, \ \alpha \leqslant \gamma < \infty)$$

that is as the limit of the mean value corresponding to a finite interval in the process when the length of the interval is increased unlimitedly. It is easy to verify that if the limit exists it does not depend on the choice of the value γ (which can also be taken as equal to α).

Let us consider an alternating current circuit in which the current flow j and the voltage u are expressed by the formulas $j = j_0 \cos(\omega t + \alpha)$ and $u = u_0 \cos(\omega t + \alpha + \varphi)$ where φ is a constant phase shift between the voltage and the current. The mean power of the current in this circuit is equal to

$$\bar{h} = \overline{ju} = \lim_{T \to \infty} \frac{1}{T} \int_{0}^{T} j_0 \cos(\omega t + \alpha)\, u_0 \cos(\omega t + \alpha + \varphi)\, dt =$$

$$= \lim_{T \to \infty} \frac{j_0 u_0}{2T} \int_{0}^{T} [\cos(2\omega t + 2\alpha + \varphi) + \cos \varphi]\, dt =$$

$$= \lim_{T \to \infty} \left\{ \frac{j_0 u_0}{4\omega} \frac{\sin(2\omega T + 2\alpha + \varphi) - \sin(2\alpha + \varphi)}{T} + \frac{j_0 u_0}{2} \cos \varphi \right\} = \frac{j_0 u_0}{2} \cos \varphi$$

This formula accounts for the significance of the quantity $\cos \varphi$ in electrical engineering.

In conclusion, we give one more inequality which is sometimes applied. Since the absolute value of a sum cannot exceed the sum of the absolute values (see the end of Sec. I.5) we can write the inequality

$$\left| \sum_{k=1}^{n} f(\xi_k)\, \Delta x_k \right| \leqslant \sum_{k=1}^{n} |f(\xi_k)\, \Delta x_k| = \sum_{k=1}^{n} |f(\xi_k)|\, \Delta x_k$$

for integral sum (6). Then passing to the limit we get

$$\left| \int_{a}^{b} f(x)\, dx \right| \leqslant \int_{a}^{b} |f(x)|\, dx \tag{20}$$

In other words, the absolute value of an integral does not exceed the integral of the absolute value of the integrand [think in what cases inequality (20) turns into the equality].

§ 2. Applications of Definite Integral

6. Two Schemes of Application. There are two basic schemes of application of the definite integral to calculating geometrical, physical and other quantities.

The first scheme is based on the definition of the integral which introduces it as the limit of an integral sum [see formula (7)]. According to this scheme, a quantity in question is approximately represented as an integral sum so that the representation should become more and still more precise, as the lengths of the subintervals of partitions are decreased, and quite exact after the passage to the limit. Therefore the quantity turns out to be equal to the limit of the integral sums, i.e. to the integral. This method was clearly illustrated by the examples considered in Sec. 1 which led to four integrals (8). As it was indicated in Sec. 2, this scheme is based on the representation of a definite integral in the form of a sum of infinitely many infinitesimal summands.

The second scheme of application of integrals consists in forming a relationship between the differentials of quantities in question, that is in forming a so-called **differential equation**. After the relationship between the differentials has been deduced we apply formula (13), which can also be put down in the form

$$\int dy = y_{terminal} - y_{initial}$$

and thus obtain a relationship between the quantities themselves. The meaning of the above formula is that the sum of infinitesimal increments of a quantity is equal to the total increment of the quantity.

Let us consider an example. Suppose a point is moving along the s-axis under the action of a variable force which is directed along the axis and assumes the value $F(s)$ at each point s. Let the point pass the distance from $s = a$ to $s = b$ and let it be necessary to compute the work A_{total} of the force along this path. The work A of the force performed in the process of motion is connected by a functional relationship with the distance passed, that is $A = A(s)$. When the point passes a small interval from s to $s + \Delta s$ the force does not change considerably and we can therefore approximately regard it as constant along this small path and, according to the well-known physical formula, we can write

$$\Delta A \approx F(s)\, \Delta s$$

A more precise formula has the form

$$\Delta A = F\,(s)\,\Delta s + \alpha \tag{21}$$

where $|\,\alpha\,| \ll \Delta s$, i.e. α is of higher order of smallness relative to Δs. The fact that α is indeed an infinitesimal of higher order of smallness is implied by the following consideration: α is caused by the variability of F on the interval Δs but the variation of F is infinitesimal when Δs is infinitesimal and, besides, this variation is multiplied by Δs when ΔA is calculated.

Now if we recall that a differential is defined as the principal (linear) part of the corresponding increment (see Sec. IV.8) we can write, on the basis of (21), that

$$dA = F\,(s)\,ds \tag{22}$$

Integrating we obtain

$$A_{total} = A\,(b) - A\,(a) = \int dA = \int_a^b F\,(s)\,ds$$

This formula is often written in a simplified form as

$$A = \int F\,ds$$

Although the limits of integration are not put down in the last formula the integral is understood as a definite integral having certain limits of integration.

In problem-solving practice the above detailed consideration is usually replaced by the following simplified consideration: the force can be regarded as constant along the infinitesimal path ds and this immediately implies formula (22) for the corresponding infinitesimal increment of the work. Then formula (22) is integrated etc. This consideration is brief but quite correct, and if we discuss it at length we shall arrive at the comprehensive consideration which was given previously. We shall turn back to this question in Sec. XVI.4.

7. **Differential Equation with Variables Separable.** The general form of a relationship of type (22) can be written as

$$dy = f\,(x)\,dx \tag{23}$$

where x and y are some variables connected by a functional relationship. Integrating we deduce

$$y_1 - y_0 = \int_{x_0}^{x_1} f\,(x)\,dx$$

where $y_0 = y\,(x_0)$ and $y_1 = y\,(x_1)$.

Equation (23) is the simplest differential equation. Differential equations will be treated in more detail in Chapter XV but some simple examples that can be considered without applying the general theory can be illustrated here. For instance, we often encounter differential equations of the form

$$dy = \varphi(y)\, dx \qquad (24)$$

We cannot simply integrate both sides of the equation because in this case the integrand under the sign of integration on the right-hand side would contain an unknown function $y(x)$. We must therefore transpose $\varphi(y)$ to the left-hand side, that is we must write $\frac{dy}{\varphi(y)} = dx$ beforehand. Then the integration yields

$$\int_{y_0}^{y_1} \frac{dy}{\varphi(y)} = x_1 - x_0 \qquad (y_0 = y(x_0), \quad y_1 = y(x_1))$$

Similarly, an equation of the form

$$dy = f(x)\, \varphi(y)\, dx \qquad (25)$$

is integrated as follows:

$$\frac{dy}{\varphi(y)} = f(x)\, dx, \quad \int_{y_0}^{y_1} \frac{dy}{\varphi(y)} = \int_{x_0}^{x_1} f(x)\, dx$$

Equations (23)-(25) are called **differential equations with variables separable** because the terms containing x and dx can be separated from the terms containing y and dy by means of simple algebraic transformations, and after this the integration is carried out immediately.

As an example, let us consider the problem of outflow of a liquid from a cylindric vessel through an opening of area σ at the bottom of the vessel (see Fig. 251). Here the height h of the level of the liquid above the bottom depends on the time t, i.e. $h = h(t)$. If the liquid is not viscous, and if it is permissible to neglect the forces of surface tension, the exit velocity v with which the liquid flows out of the vessel is described, within a sufficient accuracy, by Torricelli's law (established by E. Torricelli, 1608-1647, a prominent Italian physicist and mathematician):

$$v = \sqrt{2gh} \qquad (26)$$

We can readily form the differential equation of the problem on the basis of this law. Let us involve brief considerations similar to that in the last paragraph of Sec. 6. The exit velocity can be regarded as constant during the time interval dt, and therefore, by

formula (26), the corresponding outflow is the volume $dV = \sigma \cdot v\, dt = \sigma \sqrt{2gh}\, dt$ of the liquid.

On the other hand, the same volume is equal to $dV = S \mid dh \mid = -S\, dh$. (One should take into account that h decreases here and therefore $dh < 0$.) Equating both expressions of the volume we obtain the equation

$$-S\, dh = \sigma \sqrt{2gh}\, dt \qquad (27)$$

belonging to type (24). In order to integrate the equation let us separate the terms depending on h (and on dh) from the terms depending on t (and dt):

$$-\frac{S\, dh}{\sigma \sqrt{2gh}} = dt$$

Integrating we receive

$$-\int_{H}^{0} \frac{S\, dh}{\sigma \sqrt{2gh}} = T, \qquad (H = h(0))$$

where T is the total time of outflow of the liquid. We finally obtain

$$-\frac{S}{\sigma \sqrt{2g}} 2\sqrt{h}\Big|_{h=H}^{h=0} = T, \quad \text{i.e.} \quad T = \frac{S}{\sigma} \sqrt{\frac{2H}{g}}$$

8. Computing Areas of Plane Geometric Figures. The application of the definite integral to computing the area of a curvilinear tra-
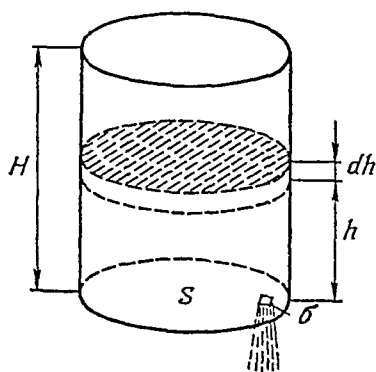
Fig. 251
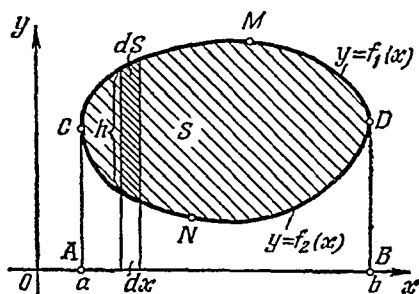
Fig. 252

pezoid was considered in Sec. 2, and the corresponding rule of signs was illustrated in Fig. 245.

If it is necessary to compute the whole area shaded in Fig. 245 in the "arithmetic" sense (i.e. not in the "algebraic sense") we can

utilize the formula

$$S_1 + S_2 + S_3 = \int_a^b |f(x)| \, dx$$

where the last integral can be evaluated according to the scheme described in example 1 of Sec. 4.

The computation of the areas of figures other than curvilinear trapezoids can also be performed with the help of integrals. For instance, the area of the geometric figure shown in Fig. 252 can be obtained as the difference of the areas of the two curvilinear trapezoids $ACMDBA$ and $ACNDBA$, i.e.

$$S = \int_a^b f_1(x) \, dx - \int_a^b f_2(x) \, dx = \int_a^b [f_1(x) - f_2(x)] \, dx = \int_a^b h(x) \, dx \quad (28)$$

where $h(x)$ is the length of the segment which is formed when the straight line parallel to the y-axis and passing through the point x of the x-axis crosses the figure.

Formula (28) can also be interpreted as follows. The area of the portion of our figure lying to the left of the straight line $x = \mathrm{const}$ depends on x. If we give x an infinitesimal increment dx (see Fig. 252) then the area of the strip shown in Fig. 252 is added to the former area. This additional area can be regarded, to within infinitesimals of higher order of smallness, as a rectangle [compare this with the deduction of formulas (21) and (22)]. It follows that $dS = h(x) \, dx$. Now, integrating, we obtain formula (28) again.

The contour of a figure, that is the curve which bounds its area, is often represented in parametric form. In such cases it is advisable to perform a change of variables in the integrals in question and choose the parameter as a new variable.

For example, let us compute the area bounded by the x-axis and by an arc of a cycloid (see Sec. II.6) with parametric equations (II.12). We mean here a part of the cycloid which connects two neighbouring points of intersection of the cycloid with the x-axis (i.e. an arc lying between two cusps), and the parameter should therefore be taken within the limits $0 \leqslant \psi \leqslant 2\pi$. Hence,

$$S = \int_0^{2\pi R} y \, dx = \int_0^{2\pi} R(1 - \cos \psi) \, d[R(\psi - \sin \psi)] =$$

$$= R^2 \int_0^{2\pi} (1 - \cos \psi)^2 \, d\psi = 3\pi R^2$$

because

$$\int (1 - \cos \psi)^2 \, d\psi = \int (1 - 2 \cos \psi + \cos^2 \psi) \, d\psi = \psi - 2 \sin \psi +$$

$$+ \int \frac{1 + \cos 2\psi}{2} \, d\psi = \frac{3}{2} \psi - 2 \sin \psi + \frac{\sin 2\psi}{4} + C$$

Now let us consider the area of a geometric figure bounded by a closed contour (L) represented by its parametric equations $x = x(t)$ and $y = y(t)$. Suppose that the variable point (x, y) [where $x = x(t)$ and $y = y(t)$] describes the contour in the positive direction once when the parameter t varies from α to γ (the positive direction is understood as counterclockwise; see Fig. 253). Then

$$S = \int_a^b y_1 \, dx - \int_a^b y_2 \, dx$$

But the first integral is equal to $\int_\gamma^\beta y\dot{x} \, dt$ because x varies from a to b as t varies from γ to β (see Fig. 253) and we have $y = y_1$ and $\dot{x} \, dt = dx$ here. The second integral is transformed similarly and thus we obtain

$$S = \int_\gamma^\beta y\dot{x} \, dt - \int_\alpha^\beta y\dot{x} \, dt = - \int_\alpha^\beta y\dot{x} \, dt - \int_\beta^\gamma y\dot{x} \, dt = - \int_\alpha^\gamma y\dot{x} \, dt \quad (29)$$

Property 10 in Sec. 4 implies that the values $t = \alpha$ and $t = \gamma$ are not necessarily such that the corresponding points (x, y) should coincide with the extreme left point of the contour; the necessary condition is that the contour should be described exactly once as t varies from α to γ.

Similarly, projecting the contour on the y-axis we can deduce the formula

$$S = \int_\alpha^\gamma x\dot{y} \, dt \quad (30)$$

If we add together formulas (29) and (30) we arrive at the formula

$$S = \frac{1}{2} \int_\alpha^\gamma (x\dot{y} - y\dot{x}) \, dt \quad (31)$$

If the contour is described in the negative direction, as t increases, we must change the signs in all formulas (29)-(31).

For example, the area of an ellipse having parametric equations (II.26) can be computed on the basis of formula (31):

$$S = \frac{1}{2} \int\limits_0^{2\pi} [a \cos t \cdot b \cos t - b \sin t\, (-a \sin t)]\, dt = \frac{1}{2} \int\limits_0^{2\pi} ab\, dt = \pi ab$$

Let us proceed to compute areas in polar coordinates. Let a curve be represented by its polar equation $\rho = f(\varphi)$ and let it be necessary to compute the area of the curvilinear "sector" $\alpha \leqslant \varphi \leqslant \beta$ (see Fig. 254). If the angle φ is increased by $d\varphi$ then the area of the portion of the sector lying below the ray ON also gains an increment,
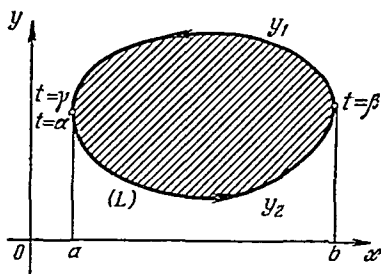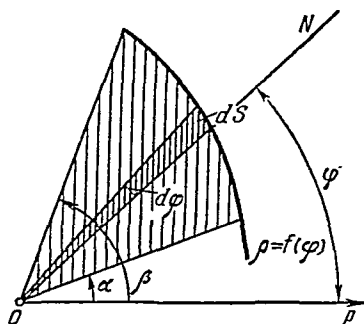
Fig. 253   Fig. 254

the additional area being regarded as an isosceles triangle with altitude ρ and base $\rho\, d\varphi$ within an accuracy to infinitesimals of higher order of smallness (why is it so?). Hence,

$$dS = \frac{1}{2} \rho\rho\, d\varphi, \quad S = \frac{1}{2} \int\limits_\alpha^\beta \rho^2\, d\varphi \qquad (32)$$

As an example, let us compute the area shaded in Fig. 255. Passing to polar coordinates in the equation of the hyperbola we obtain

$$\rho^2 \cos^2 \varphi - \rho^2 \sin^2 \varphi = 1, \quad \text{i.e.} \quad \rho^2 = \frac{1}{\cos^2 \varphi - \sin^2 \varphi}$$

Consequently, by formula (32), we derive

$$S = \frac{1}{2} \int\limits_0^\varphi \frac{1}{\cos^2 \varphi - \sin^2 \varphi}\, d\varphi = \frac{1}{4} \ln \frac{1 + \tan \varphi}{1 - \tan \varphi}$$

(verify the calculations!).

The above result implies an interesting consequence. We have

$$\frac{1+\tan\varphi}{1-\tan\varphi}=e^{4S}, \quad \text{i.e.} \quad \tan\varphi=\frac{e^{4S}-1}{e^{4S}+1}=\frac{e^{2S}-e^{-2S}}{e^{2S}+e^{-2S}}=\tanh 2S$$

(see Sec. I.28), and therefore

$$NM = \rho\sin\varphi = \frac{\sin\varphi}{\sqrt{\cos^2\varphi-\sin^2\varphi}} = \frac{\tan\varphi}{\sqrt{1-\tan^2\varphi}} =$$

$$= \frac{\tanh 2S}{\sqrt{1-\tanh^2 2S}} = \frac{\tanh 2S}{\dfrac{1}{\cosh 2S}} = \sinh 2S$$

We similarly find that $ON = \rho\cos\varphi = \cosh 2S$ and $AP = = \tan\varphi = \tanh 2S$. The comparison of these results with Fig. 256 where $\varphi = 2S$, $MN = \sin 2S$, $ON = \cos 2S$ and $AP = \tan 2S$
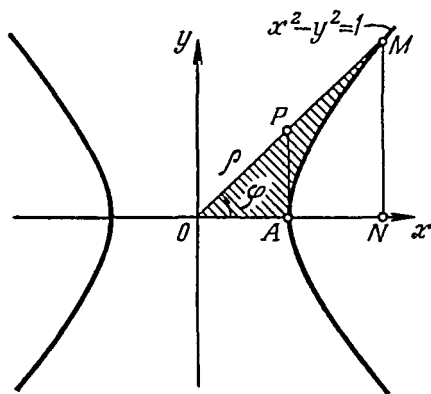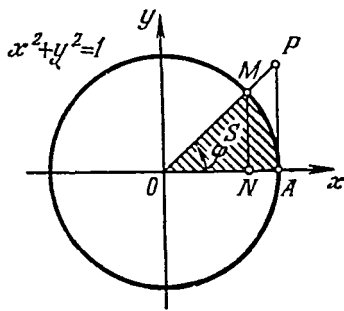
Fig. 255   Fig. 256

reveals the geometric meaning of the resemblance between trigonometric (circular) functions and hyperbolic functions and accounts for the origin of the terms "hyperbolic" sine, cosine and tangent.

9. The Arc Length of a Curve. We have already dealt with the differential of the arc length in our course (see Sec. VII.23). We shall denote it by dL:

$$dL = \sqrt{dx^2 + dy^2 + dz^2} = \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\, dt$$

We shall agree that $dL > 0$ and therefore take the sign $+$ in front of the radical. It follows that if the values $t = \alpha$ and $t = \beta$ of the parameter correspond to the end-points of the arc then its length is

$$L = \int_{\alpha}^{\beta} \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\, dt$$

For a plane curve the formula of the arc length is simplified:

$$L = \int_\alpha^\beta \sqrt{dx^2 + dy^2} = \int_\alpha^\beta \sqrt{\dot{x}^2 + \dot{y}^2}\, dt$$

If a curve is represented by an equation of the form $y = f(x)$ $(a \leqslant x \leqslant b)$ then

$$L = \int_\alpha^\beta \sqrt{dx^2 + dy^2} = \int_a^b \sqrt{1 + y'^2}\, dx$$

For instance, the arc length of the part of a cycloid [represented by equations (II.12)] between its neighbouring spinodes (cusps) is computed with the help of the above formula: $L = \int_\alpha^\beta \sqrt{\dot{x}^2 + \dot{y}^2}\, dt = $

$= \int_0^{2\pi} \sqrt{\left(\frac{dx}{d\psi}\right)^2 + \left(\frac{dy}{d\psi}\right)^2}\, d\psi$. In computing the arc length we take into account the symmetry of the arc:

$$L = 2 \int_0^\pi R \sqrt{(\psi - \sin\psi)'^2 + (1 - \cos\psi)'^2}\, d\psi =$$

$$= 2R \int_0^\pi \sqrt{(1 - \cos\psi)^2 + \sin^2\psi}\, d\psi = 2R \int_0^\pi \sqrt{2 - 2\cos\psi}\, d\psi =$$

$$= 4R \int_0^\pi \sin\frac{\psi}{2}\, d\psi = -8R \cos\frac{\psi}{2}\Big|_0^\pi = 8R$$

The result is extremely simple!

The differential of the arc length in polar coordinates is readily obtained from Fig. 257:

$$dL = \sqrt{(d\rho)^2 + (\rho\, d\varphi)^2} \tag{33}$$

It follows that if the equation of a curve in polar coordinates is given in the form $\rho = f(\varphi)$ then its arc length corresponding to the interval $\alpha \leqslant \varphi \leqslant \beta$ of variation of φ is equal to

$$L = \int_\alpha^\beta \sqrt{d\rho^2 + \rho^2\, d\varphi^2} = \int_\alpha^\beta \sqrt{\left(\frac{d\rho}{d\varphi}\right)^2 + \rho^2}\, d\varphi$$

We suggest that the reader should check that expression (33) can also be obtained from the formulas

$$dL = \sqrt{dx^2 + dy^2}, \quad x = \rho \cos \varphi \quad \text{and} \quad y = \rho \sin \varphi$$

As an example, let us consider the arc length of the cardioid depicted in Fig. 72. Taking advantage of its polar equation put down in Fig. 72 we find the arc length:

$$L = 2 \int_0^\pi \sqrt{4a^2 \sin^2 \varphi + 4a^2 (1 - \cos \varphi)^2} \, d\varphi = 16a$$

(verify the result!).

10. Computing Volumes of Solids. Suppose that we have a solid and that we know the areas of its parallel sections by planes perpendicular to an axis (see Fig. 258). Let it be necessary to compute the volume of the solid. Let x be the coordinate reckoned along the
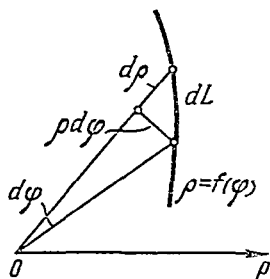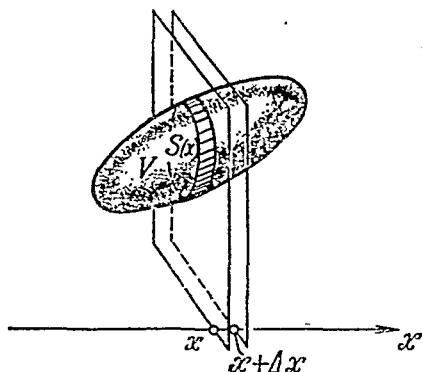
Fig. 257   Fig. 258

axis and let the area of the section by the plane passing through the point x be $S = S(x)$. Take the volume of the part of the solid lying to the left of the plane passing through x. If we increase x by $\Delta x = dx$ (for definiteness, we take $\Delta x > 0$), then the plane is moved to the right, and an additional volume is added to the former volume. This additional volume is the volume of the "slice" which can be regarded as a cylinder with a wide base of area $S(x)$ and small altitude dx, within an accuracy to infinitesimals of higher order of smallness. It follows that

$$\Delta V = S(x) \, \Delta x + \textit{infinitesimals of higher order}$$

i.e.

$$dV = S(x) \, dx$$

Now if x varies from a to b we obtain

$$V = \int\limits_{a}^{b} S(x)\, dx \qquad (34)$$

Let us take an example. We shall compute the volume of a solid which is bounded by the surface of a right circular cylinder, by the half-disc lying in a plane perpendicular to the axis of the cylinder and by the part of an oblique plane passing through the diameter of the disc. The solid is depicted in Fig. 259; it has the form of a
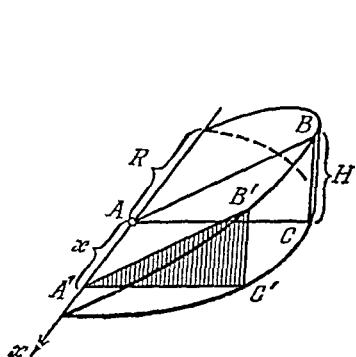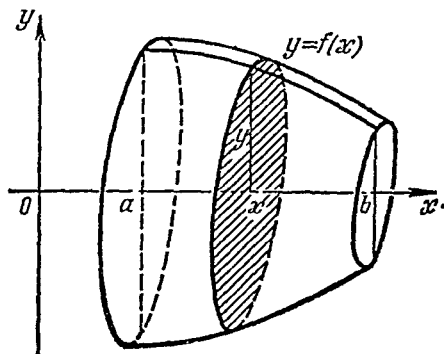
Fig 259   Fig. 260

"hoof". Since the triangles ABC and $A'B'C'$ are similar according to Fig. 259, the area of the section which is shaded in Fig. 259 is equal to

$$S = \frac{1}{2} RH \frac{R^2 - x^2}{R^2} = \frac{R^2 - x^2}{2R} H$$

Consequently, by formula (34), we have

$$V = 2 \int\limits_{0}^{R} S\, dx = 2 \int\limits_{0}^{R} \frac{R^2 - x^2}{2R} H\, dx = \frac{H}{R} \left(R^2 x - \frac{x^3}{3} \right) \Big|_{0}^{R} = \frac{2}{3} R^2 H$$

Note that the number π does not enter into the answer!

We now consider the volume of a solid of revolution. Let a curve having the equation $y = f(x)$ rotate in space about the x-axis and let it describe the boundary surface of the solid of revolution. Then the area of the section of the solid by the plane perpendicular to the x-axis and passing through the point x is equal to $S = \pi y^2$ where $y = f(x)$ (see Fig. 260). Hence, by formula (34), we have

$$V = \pi \int\limits_{a}^{b} y^2\, dx = \pi \int\limits_{a}^{b} f^2(x)\, dx \qquad (35)$$

For instance, we can regard a sphere of radius R as a solid generated by the revolution of the semi-circle having the equation $y =$

$= \sqrt{R^2 - x^2}$. The volume of the sphere is therefore equal to

$$V = \pi \cdot 2 \int_0^R (\sqrt{R^2 - x^2})^2 \, dx = 2\pi \left(R^2 x - \frac{x^3}{3} \right) \Big|_0^R = \frac{4}{3} \pi R^3$$

Compare the above deduction of the formula with the long and artificial procedure applied to deducing the formula of the volume of a sphere in elementary mathematical courses.

11. Computing Area of Surface of Revolution. The formula for computing the area of an arbitrary surface will be established in
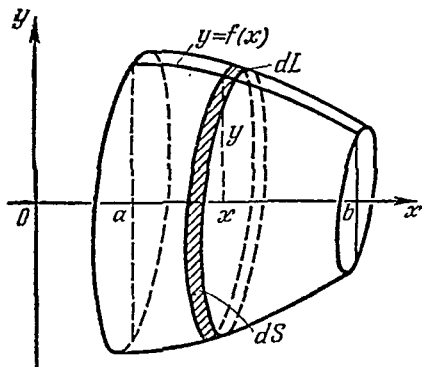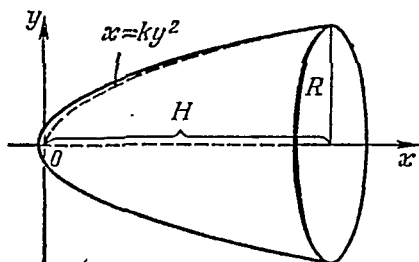
Fig. 261   Fig. 262

Sec. XVI.10. But the computation of the area of a surface of revolution can be discussed now. Let a curve $y = f(x) > 0$ rotate about the x-axis (see Fig. 261). Let us draw a plane perpendicular to the x-axis and passing through the point x where x is considered to be variable. Take the area of the portion of the surface of revolution lying to the left of the plane.

If the plane is translated by the distance dx to the right the area of the portion is increased by the surface element which is shaded in Fig. 261. The element has the form of a ring-shaped surface of width dL with the circumference (length) $2\pi y$ because the radius of the ring is equal to y. Consequently, we have

$$dS = 2\pi y \, dL = 2\pi y \sqrt{1 + y'^2} \, dx$$

and thus

$$S = 2\pi \int_a^b y \, dL = 2\pi \int_a^b y \sqrt{1 + y'^2} \, dx \qquad (36)$$

As an example, let us compute the area of the portion of a paraboloid of revolution intercepted by a plane perpendicular to the axis of revolution (see Fig. 262). Let the radius R of the base and the altitude H be given. The surface being generated by the revolution of a parabola whose axis coincides with the x-axis, the equation of the curve has the form $x = ky^2$. The constant k must be chosen so that the parabola should pass through the point (H, R). This implies $H = kR^2$, i.e. $k = \frac{H}{R^2}$. Finally, the equation of the curve is

$$x = \frac{H}{R^2} y^2 \quad [\text{or} \quad y = R \sqrt{\frac{x}{H}}$$

Taking advantage of formula (36) we obtain

$$S = 2\pi \int_0^h R \sqrt{\frac{x}{H}} \sqrt{1 + \left[\left(R\sqrt{\frac{x}{H}}\right)'\right]^2} \, dx =$$

$$= 2\pi \frac{R}{\sqrt{H}} \int_0^H \sqrt{x} \sqrt{1 + \frac{R^2}{H} \cdot \frac{1}{4x}} \, dx = \pi \frac{R}{H} \int_0^H \sqrt{4xH + R^2} \, dx =$$

$$= \pi \frac{R}{H} \frac{(4xH + R^2)^{3/2}}{\left(\frac{3}{2}\right) \cdot 4H} \Bigg|_0^H = \frac{\pi R}{6H^2} [(4H^2 + R^2)^{3/2} - R^\circ]$$

§ 3. Numerical Integration

12. General Remarks. The basic method of evaluating a definite integral by means of the corresponding indefinite integral described in Sec. 3 is sometimes inexpedient and even practically inapplicable. As it was indicated in Sec. XIII.11, there are many indefinite integrals of elementary functions which cannot be expressed in terms of elementary functions or which have such expressions that are too complicated. Besides, a function which we have to integrate can be represented in a way which does not yield its analytical expression. In these cases one can use a number of methods which we are going to review here.

1. Some integrals are expressible in terms of certain thoroughly studied and tabulated non-elementary "special" functions.

For instance, one of these functions is the **error function**

$$\text{Erf } x = \int_0^x e^{-t^2} dt \quad (-\infty < x < \infty) \tag{36'}$$

Further examples are the **Fresnel integrals** (named after A. Fresnel, 1788-1827, a prominent French physicist, the creator of the wave

theory of light),

$$C(x) = \int\limits_0^x \cos \frac{\pi t^2}{2} dt \quad \text{and} \quad S(x) = \int\limits_0^x \sin \frac{\pi t^2}{2} dt \quad (-\infty < x < \infty)$$

the exponential integral

$$\text{Ei } x = \int\limits_{-\infty}^x \frac{e^t}{t} dt \quad (-\infty < x < 0)$$

the sine integral

$$\text{Si } x = \int\limits_0^x \frac{\sin t}{t} dt \quad (-\infty < x < \infty)$$

the cosine integral

$$\text{Ci } x = \int\limits_\infty^x \frac{\cos t}{t} dt \quad (0 < x < \infty)$$

and many other functions. Integrals with infinite limits will be considered in full in § 4.

Let us take an example. In order to evaluate the integral

$$I = \int\limits_0^1 \frac{\sin^2 x}{x^2} dx$$

we integrate it by parts putting $u = \sin^2 x$ and $dv = x^{-2} dx$:

$$I = -\frac{\sin^2 x}{x} \Big|_0^1 + \int\limits_0^1 \frac{2 \sin x \cos x}{x} dx = -\sin^2 1 + \int\limits_0^1 \frac{\sin 2x}{x} dx$$

Now performing the change of variable $2x = t$ we get

$$I = -\sin^2 1 + \int\limits_0^2 \frac{\sin t}{t} dt = -\sin^2 1 + \text{Si } 2 = 0.8973$$

The value of the sine integral is taken from [23]. A great number of special functions are described in this book.

2. It is sometimes possible to find the exact value of a definite integral with certain specific limits without calculating the corresponding indefinite integral. Calculations of this type are usually difficult but nevertheless we shall give some examples further [for instance, see formula (72)]. Many integrals of this kind are collected in [19].

For example, in this book we can find the formulas

$$\int\limits_{0}^{\frac{\pi}{2}} \tan^p x \, dx = \frac{\pi}{2} \left(\cos \frac{p\pi}{2} \right)^{-1} \quad (-1 < p < 1),$$

$$\int\limits_{0}^{\pi} \ln \sin x \, dx = -\pi \ln 2 \quad \text{etc.}$$

but the corresponding indefinite integrals are not elementary functions.

3. Expansions of the integrand into series of different types are also often used for integration. This method will be described at length in Chapter XVII but the simple power series which were mentioned in Sec. IV.16 can be readily applied now.

For instance, taking series (IV.55) for the function e^x we obtain

$$\int\limits_{0}^{1} \frac{e^x - 1}{x} \, dx = \int\limits_{0}^{1} \frac{1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots - 1}{x} \, dx =$$

$$= \int\limits_{0}^{1} \left(\frac{1}{1!} + \frac{x}{2!} + \frac{x^2}{3!} + \ldots \right) dx = \frac{1}{1!} + \frac{1}{2 \cdot 2!} + \frac{1}{3 \cdot 3!} + \ldots = 1.318$$

(the result is accurate to 0.001).

It was noted in Sec. IV.16 that in practice such series can be treated as finite sums in which the number of terms is taken depending on the accuracy chosen.

4. Graphical integration is used when a function is represented by its graph. The method is based on the geometric meaning of the definite integral (see Sec. 2) which implies that the integral is equal to the area of the corresponding curvilinear trapezoid. We can compute the area either by constructing the graph on the plotting paper and figuring the number of squares lying inside the bounding line or by using a special instrument, the so-called *planimeter*. After the tracer of the planimeter has been passed round the periphery of the area of an arbitrary form which is to be measured we read the area on the meter of the planimeter. Since a planimeter is a simple mechanism integration with its help is called mechanical integration.

5. The most comprehensive methods applicable to integrals of arbitrary functions are the methods of numerical integration which are reviewed in Sec. 13. These methods can be used for functions represented in any possible way, especially for functions represented by means of tables.

13. Formulas of Numerical Integration. These formulas make it possible to evaluate approximate values of a definite integral if

the values of the integrand at certain points (so-called *nodes*) of the interval of integration are given.

Let us begin with the most elementary formula. Let it be necessary to evaluate the integral

$$\int_a^b y\,dx, \quad y = f(x) \tag{37}$$

Suppose that the interval of integration $a \leqslant x \leqslant b$ is divided into a finite number n of equal parts and that the values of the integrand at the points of division are given or calculated.

Introduce the notation

$$\frac{b-a}{n} = h, \quad f(a) = y_0, \quad f(a+h) = y_1,$$

$$f(a+2h) = y_2, \quad \ldots, \quad f(a+nh) = f(b) = y_n$$

If we draw the ordinates at each of the nodes the curvilinear trapezoid whose area is equal to integral (37) is broken into n parts (see Fig. 263). Each of these parts is also a curvilinear trapezoid. Now let us replace the parts by rectilinear trapezoids whose bases are pairs of neighbouring ordinates (see Fig. 263). The areas of these trapezoids are equal to

Fig. 263

$$\frac{y_0+y_1}{2}h, \quad \frac{y_1+y_2}{2}h, \quad \ldots, \quad \frac{y_{n-1}+y_n}{2}h$$

Adding the areas together we obtain the area of a polygonal figure inscribed into the original curvilinear trapezoid. If n is sufficiently large, that is if h is sufficiently small, the area of the polygon will be approximately equal to the area of the curvilinear trapezoid, i.e. to the integral. Thus, we obtain

$$\int_a^b y\,dx \approx \frac{y_0+y_1}{2}h + \frac{y_1+y_2}{2}h + \ldots + \frac{y_{n-1}+y_n}{2}h$$

or

$$\int_a^b y\,dx \approx h\left(\frac{y_0+y_n}{2} + y_1 + y_2 + \ldots + y_{n-1}\right) \tag{38}$$

This is the so-called **trapezoid formula** (trapezoid rule).

We can give an interpretation of the trapezoid formula which is independent of its geometric meaning. Virtually, before inte-

grating the function we substituted it by linear functions which take on the values of the integrand $y = f(x)$ assumed at the endpoints of each interval of the form $a \leqslant x \leqslant a + h$, $a + h \leqslant x \leqslant a + 2h$ etc. Hence, we can say that we have performed linear interpolation (see Sec. I.22) before evaluating the integral. Now let us recall that we gave interpolation formulas in § V.2 which approximate functions with a greater accuracy than formulas of linear interpolation. Therefore formulas of numerical integration based on these interpolation formulas are much more accurate than formula (38).
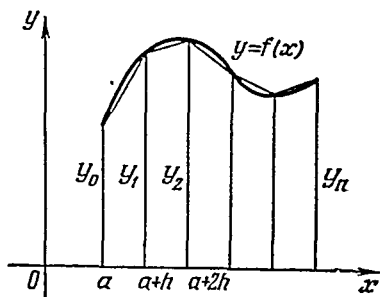
If we use interpolation polynomials of the second degree we shall arrive at **Simpson's formula** named after the English mathematician T. Simpson (1710-1761) who deduced the formula. Let us first suppose that we know the values of the integrand at three points of the form x_0, $x_0 + h$ and $x_0 + 2h$, that is we are given

$$y(x_0) = y_0, \quad y(x_0 + h) = y_1 \quad \text{and} \quad y(x_0 + 2h) = y_2$$

Then we can put down the interpolation polynomial of the second degree which assumes the same values at these points. By Newton's formula (see Sec. V.8), the polynomial is

$$P(x) = y_0 + \Delta y_0 \frac{s}{h} + \frac{\Delta^2 y_0}{2} \frac{s}{h} \left(\frac{s}{h} - 1 \right) \quad (s = x - x_0)$$

Now performing the interpolation on the subinterval $x_0 \leqslant x \leqslant x_0 + 2h$ we get, by means of the substitution $x - x_0 = s$, $dx = ds$, $0 \leqslant s \leqslant 2h$, the expression

$$\int_{x_0}^{x_0 + 2h} P(x)\, dx = \int_0^{2h} \left[y_0 + \Delta y \frac{s}{h} + \frac{\Delta^2 y_0}{2} \left(\frac{s^2}{h^2} - \frac{s}{h} \right) \right] ds =$$

$$= y_0 \cdot 2h + \Delta y_0 \cdot 2h + \frac{\Delta^2 y_0}{2} \cdot \frac{2}{3} h$$

Further, if we substitute

$$\Delta y_0 = y_1 - y_0,$$

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = (y_2 - y_1) - (y_1 - y_0) = y_2 - 2y_1 + y_0$$

then after collecting similar terms we obtain

$$\int_{x_0}^{x_0 + 2h} f(x)\, dx \approx \int_{x_0}^{x_0 + 2h} P(x)\, dx =$$

$$= 2h \left[y_0 + (y_1 - y_0) + \frac{1}{6}(y_2 - 2y_1 + y_0) \right] = h\, \frac{y_0 + 4y_1 + y_2}{3} \quad (39)$$

Now suppose that the interval of integration $a \leqslant x \leqslant b$ is divided into $2n$ equal subintervals with the help of the points of division

$$x_0 = a, \quad x_1 = a + h, \quad x_2 = a + 2h, \ldots$$
$$\ldots, x_{2n} = a + 2nh = b$$

where $h = \frac{b-a}{2n}$. Then we can apply formula (39) to each pair of subintervals of the form $x_0 + (2k - 2) h \leqslant x \leqslant x_0 + (2k - 1) h$ and $x_0 + (2k - 1) h \leqslant x \leqslant x_0 + 2kh$ $(k = 1, 2, \ldots, n)$:

$$\int_{x_0}^{x_0+2h} y\, dx \approx h\, \frac{y_0+4y_1+y_2}{3}, \qquad \int_{x_0+2h}^{x_0+4h} y\, dx \approx h\, \frac{y_2+4y_3+y_4}{3}, \quad \ldots$$

$$\ldots, \int_{x_0+(2n-2)h}^{x_0+2nh} y\, dx \approx h\, \frac{y_{2n-2}+4y_{2n-1}+y_{2n}}{3}$$

Adding together these formulas and collecting similar terms we receive Simpson's formula:

$$\int_a^b y\, dx \approx \frac{h}{3}\, [(y_0+y_{2n})+2\,(y_2+y_4+\ldots+y_{2n-2})+4\,(y_1+y_3+\ldots+y_{2n-1})]$$

$$\tag{40}$$

Now we proceed to estimate the accuracy of formulas (38) and (40). Newton's formula (V.27) implies that in performing linear interpolation on a subinterval we get an error of the order of $\Delta^2 y$, i.e. of the order of h^2 (see Sec. V.7). According to formula (18), we can estimate the corresponding absolute error of the integral taken over the subinterval if we multiply the error of the interpolation by the length h of the subinterval of integration. Hence, the error of the integral on a subinterval is of the order of h^3. Formula (38) is obtained by adding together n approximate formulas having the errors of the order of h^3. Therefore, the number of subintervals being equal to $n = \frac{b-a}{h}$, the resultant error is of the order of

$$n \cdot h^3 = \frac{b-a}{h}\, h^3 = (b-a) \cdot h^2$$

that is of the order of h^2. For instance, if we increase the number of the division points twice the degree of accuracy of formula (38) will increase, approximately, four times.

One can think that analogous considerations applied to formula (40) must indicate that its error is of the order of h^3. But this is wrong because in reality the accuracy is still higher. Actually, when dropping the term containing $\Delta^3 y_0$ which enters into Newton's formula we make an error of the order of h^3. But it turns out that the integral of the term is equal to zero because

$$\int_{x_0}^{x_0+2h} \frac{s}{h} \left(\frac{s}{h}-1\right) \left(\frac{s}{h}-2\right) dx = \int_0^{2h} \left[\left(\frac{s}{h}\right)^3 - 3\left(\frac{s}{h}\right)^2 + 2\frac{s}{h}\right] ds =$$

$$= \left(\frac{s^4}{4h^3} - \frac{3s^3}{3h^2} + \frac{2s^2}{2h}\right) \Big|_0^{2h} = 0$$

Hence, the error which occurs, after integration, is determined by the subsequent term of Newton's formula. This subsequent term is of the order of h^4. Consequently, the error of formula (39) is of the order of h^5 and the error of final formula (40) is of the order of h^4. For example, if the number of the points of division is increased twice the accuracy of formula (40) increases 16 times. At the same time the application of formula (40) is not much more complicated than that of formula (38).

Let us take an example. The exact value of the integral $I = \int_0^1 \frac{1}{1+x^2}\, dx$ is readily found:

$$I = \int_0^1 \frac{1}{1+x^2}\, dx = \text{arc tan } x \,|_0^1 = \frac{\pi}{4} = 0.785$$

If we did not know the answer we could evaluate the integral approximately by means of formula (38) or (40). For simplicity's sake, let us take $n = 2$, that is

$$h = 0.5, \quad x_0 = 0, \quad x_1 = 0.5, \quad x_2 = 1, \quad y_0 = 1.000,$$
$$y_1 = 0.800, \quad y_2 = 0.500$$

Applying formula (38) we get the value

$$I \approx 0.5 \left(\frac{1.000 + 0.500}{2} + 0.800 \right) = 0.775$$

whose error is ≈ 1.3 per cent. The calculations according to formula (40) yield the value

$$I \approx \frac{0.5}{3} (1.000 + 0.500 + 4 \times 0.800) = 0.783$$

the error being about 0.3 per cent. If we had taken $n = 10$ the error of formula (38) would have been about 0.05 per cent and the error of formula (40) about 10^{-6} per cent.

In books devoted to numerical methods in mathematics one can find formulas of approximate integration that are more accurate than Simpson's formula. We sometimes use formulas constructed for nodes which are not equally spaced.

§ 4. Improper Integrals

Up till now we have considered definite integrals with finite intervals of integration and with integrands which do not approach infinity on the intervals. We shall call such integrals proper. If at least one of the above conditions is not fulfilled the integral is called **improper**. A proper integral of a continuous function (and

of a function of some wider class) always has a certain numerical value. In contrast to it, improper integrals which we are going to study here may not have such a value.

14. Integrals with Infinite Limits of Integration. First let us take an integral of the form

$$I = \int_a^\infty f(x)\, dx \tag{41}$$

where the lower limit a and the integrand $f(x)$ (considered on $a \leqslant$ $\leqslant x < \infty$) are supposed to be finite. This integral is improper because its upper limit is infinitely large.

To define integral (41) in an exact sense we use the same approach as that applied in Sec. III.6 to the sum of an infinite series. Namely, we first "truncate" the integral, i.e. we cut off an infinite portion of its interval of integration and consider the integral

$$\int_a^N f(x)\, dx \tag{42}$$

where N is a large but finite number. Integral (42) is proper and possesses a certain numerical value. Then we make N tend to infinity because in integral (41) we have the sign of infinity as the upper limit of integration. Integral (42) varies in a certain manner as $N \to \infty$. If, in this process, it has a certain finite limit we say that integral (41) is **convergent**. In this case we put, by definition,

$$\int_a^\infty f(x)\, dx = \lim_{N \to \infty} \int_a^N f(x)\, dx \tag{43}$$

If there is no finite limit integral (41) is said to be **divergent**. In such a case we shall not define a numerical value of the integral in our course (although even in this case it is sometimes possible to speak about the value of the integral). Hence, we shall speak about the numerical value of an improper integral of type (41) only if it converges.

Let us note a particular case of divergence: if integral (42) has an infinite limit as $N \to \infty$ then integral (41) is said *to diverge to infinity*, in this case formula (43) makes sense and can be used.

Let us consider several examples. Suppose a point T is moving under the action of a force which is directed along the straight line connecting T with a fixed point O (from T to O) and whose absolute value is inversely proportional to the square of the distance from O to T. In particular, gravitational force and force of attraction between two charges of electricity are of this kind. Suppose it is necessary to compute the work which should be expended to

remove the point T from a position T_0 into infinity. This work is called the **potential** of the force.

To perform the computation let us write the expression of the force:

$$F = \frac{k}{s^2} \qquad (s = OT)$$

where k is a proportionality factor. Then, by Sec. 6, the work is

$$A = \int_{s_0}^{\infty} \frac{k}{s^2}\, ds \qquad (s_0 = OT_0) \tag{44}$$

This is an improper integral which must be calculated by formula (43):

$$A = \lim_{N \to \infty} \int_{s_0}^{N} \frac{k}{s^2}\, ds = \lim_{N \to \infty} \left. -\frac{k}{s} \right|_{s=s_0}^{s=N} = \lim_{N \to \infty} \left(\frac{k}{s_0} - \frac{k}{N} \right) = \frac{k}{s_0}$$

Thus, integral (44) converges. We see that the potential is inversely proportional to the first power of the distance from O to T_0. At first glance one can find it strange that the work corresponding to the motion along an infinite path turns out to be finite although,
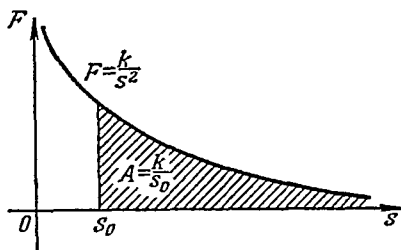
Fig. 264

theoretically, the force never stops acting on the point. The explanation is that the force decreases so fast as the point moves towards infinity that the expended work tends to a finite quantity but not to infinity although it increases all the time. The geometric meaning of the result is illustrated in Fig. 264: despite the fact that the shaded geometric figure extends to infinity its altitude $\Big($that is its ordinate $F = \frac{k}{s^2}\Big)$ decreases so fast that its area turns out to be finite.

Of course, in reality s varies within a finite range extending not to infinity but to a finite value S (which is very large) because all physical quantities are finite. Hence, in reality we have an integral

of the form

$$\int\limits_{s_0}^{S} \frac{k}{s^2}\, ds \tag{45}$$

But beginning with some sufficiently large value of S this integral does not practically vary and therefore it can be replaced by the "limiting" integral (44) because it is easier to investigate (44) in theoretical studies as the value S is not known exactly. The significance of the convergence of integral (44) is that it provides for the possibility of replacing "real" integral (45) by integral (44) for large S. From the physical point of view this means that the action of O upon T can be neglected when the distance from the point O becomes sufficiently large. In performing such a replacement we do not need the exact value of S; the only important thing here is that we must be sure of S being sufficiently large.

As a second example, let us consider the improper integral

$$\int\limits_{1}^{\infty} \frac{1}{x}\, dx \tag{46}$$

Since

$$\lim_{N\to\infty} \int\limits_{1}^{N} \frac{1}{x}\, dx = \lim_{N\to\infty} \ln x \Big|_{1}^{N} = \lim_{N\to\infty} \ln N = \ln\infty = \infty$$

we see that integral (46) diverges to infinity. Thus, we can write

$$\int\limits_{1}^{\infty} \frac{1}{x}\, dx = \infty$$

Finally, consider the improper integral

$$\int\limits_{0}^{\infty} \sin x\, dx \tag{47}$$

In this case we have neither a finite nor an infinite value of the limit

$$\lim_{N\to\infty} \int\limits_{0}^{N} \sin x\, dx = \lim_{N\to\infty} (-\cos x)\Big|_{0}^{N} = \lim_{N\to\infty} (1-\cos N)$$

because it does not exist since the values of $\cos N$, as $N\to\infty$, "oscillate" within the limits from -1 to $+1$ all the time. Consequently, integral (47) does not diverge to infinity but it diverges in an oscillating way, its values constantly varying from 0 to 2

and back from 2 to 0. Hence, the integral has neither a finite nor an infinite value.

15. Basic Properties of Integrals with Infinite Limits of Integration. Many properties of proper integrals are automatically extended to improper integrals of form (41).

First of all, if an antiderivative $F(x)$ of the function $f(x)$ is known in the interval $a \leqslant x < \infty$ then

$$\int_a^\infty f(x)\,dx = \lim_{N \to \infty} \int_a^N f(x)\,dx = \lim_{N \to \infty}[F(N) - F(a)] = F(\infty) - F(a)$$

because $F(\infty)$ is nothing but the notation for $\lim_{N \to \infty} F(N)$. Hence, in this case improper integral (41) can be evaluated by means of formula (11) deduced for proper integrals. The expression $F(\infty)$ itself indicates whether the integral diverges or converges.

For instance, in examples (44), (46) and (47) we could have calculated in the following way:

$$\int_{s_0}^\infty \frac{k}{s^2}\,ds = -\frac{k}{s}\Big|_{s=s_0}^{s=\infty} = \frac{k}{s_0} - \frac{k}{\infty} = \frac{k}{s_0},$$

$$\int_1^\infty \frac{1}{x}\,dx = \ln x\Big|_1^\infty = \ln\infty - \ln 1 = \infty \text{ and}$$

$$\int_0^\infty \sin x\,dx = -\cos x\Big|_0^\infty = -\cos\infty + 1$$

The last result shows in fact that integral (47) does not exist since the expression $\cos\infty$ makes no sense.

All the basic properties enumerated in Sec. 4 also remain true for improper integrals with natural exceptions involving the case of a divergent integral. For instance, formulas (18) and (19) no longer hold because an integral of a nonzero constant taken over an infinite interval always diverges. We also note the following simple property: if integral (41) converges we have

$$\int_N^\infty f(x)\,dx = \int_a^\infty f(x)\,dx - \int_a^N f(x)\,dx \to 0$$

If it is difficult to compute the corresponding indefinite integral we usually begin the investigation of the improper integral with testing its convergence or divergence on the basis of special tests which we are going to study now.

First of all, it should be noted that the convergence or divergence of integral (47) is determined only by the behaviour of the function $f(x)$ as x approaches infinity, that is for sufficiently large x. In other words, the integrals $\int_a^\infty f(x)\,dx$ and $\int_b^\infty f(x)\,dx$ converge or diverge simultaneously provided $f(x)$ does not approach infinity at a point lying between a and b. In fact, the difference between the integrals is a proper integral having a certain finite numerical value and therefore it cannot violate the convergence in case one of the integrals converges and it cannot provide for the convergence if one of the integrals diverges.

We first consider an integral of a non-negative function:

$$\int_a^\infty f(x)\,dx, \quad f(x) \geqslant 0 \qquad (48)$$
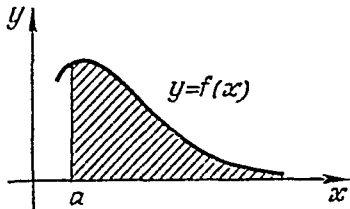
Fig. 265

Such an integral either converges or diverges to infinity since the integral taken from a to N is a non-decreasing quantity here as $N \to \infty$, and such a quantity either has a finite limit or tends to infinity (see Sec. III.5). In this case the convergence or divergence means, geometrically, that the area of the infinite plane figure shaded in Fig. 265 is, respectively, finite or infinite.

The fact that integral (48) converges or diverges in this case can be indicated, respectively, by the formula

$$\int_a^\infty f(x)\,dx < \infty$$

or

$$\int_a^\infty f(x)\,dx = \infty$$

Of course, these formulas cannot be applied to divergent integrals of "oscillating" type (47).

The simplest test for convergence is the comparison test: if

$$0 \leqslant g(x) \leqslant f(x) \quad (a \leqslant x < \infty) \qquad (49)$$

and $\int_a^\infty f(x)\,dx < \infty$, that is this integral converges, then we also

have $\int\limits_a^\infty g(x)\,dx < \infty$, that is the last integral converges too. The proof of the test is implied by integration of inequality (49) or by the geometric meaning of convergence (see Fig. 265). The same test suggests that if conditions (49) hold and the integral of $g(x)$ diverges (equals infinity) then the integral of $f(x)$ also diverges.

We also use the following test: if

$$\frac{f(x)}{g(x)} \xrightarrow[x\to\infty]{} k \neq 0 \quad (k \neq \infty) \tag{50}$$

then the integrals

$$\int\limits_a^\infty f(x)\,dx \quad \text{and} \quad \int\limits_b^\infty g(x)\,dx$$

converge or diverge simultaneously (although in the case of convergence their numerical values can considerably differ even if $k = 1$ and $a = b$). Indeed, condition (50) implies that neither of the functions $f(x)$ and $g(x)$ can be considerably larger than the other as $x \to \infty$, i.e. $f(x) \sim kg(x)$ where $\sim$ is the sign of equivalence (see Sec. III.7). Therefore, if the shaded figure of the type shown in Fig. 265 has a finite area for one of the functions the same must be true for the other.

Most often we compare a given integral of form (48) with the integral of a power function of the form

$$I = \int\limits_1^\infty \frac{1}{x^p}\,dx \tag{51}$$

which can be easily investigated in a direct manner. If $p > 1$ we have

$$I = \int\limits_1^\infty x^{-p}\,dx = \frac{x^{-p+1}}{-p+1}\Big|_1^\infty = \frac{1}{(-p+1)\,x^{p-1}}\Big|_1^\infty = \frac{1}{\infty} - \frac{1}{-p+1} = \frac{1}{p-1} < \infty$$

whereas, for $p < 1$, we have

$$I = \frac{x^{-p+1}}{-p+1}\Big|_1^\infty = \frac{x^{1-p}}{1-p}\Big|_1^\infty = \infty - \frac{1}{1-p} = \infty$$

Finally, if $p = 1$ then

$$I = \int\limits_1^\infty \frac{1}{x}\,dx = \ln x\Big|_1^\infty = \ln\infty - \ln 1 = \infty$$

Thus, integral (51) converges for $p > 1$ and diverges to infinity for $p \leqslant 1$. Hence, by test (50), we can draw the same conclusion

concerning integral (48) if

$$f(x) \sim \frac{A}{x^p} \quad \text{as} \quad x \to \infty$$

For instance,

$$\int_0^\infty \frac{dx}{\sqrt[3]{x^2+1}} = \infty$$

because

$$\frac{1}{\sqrt[3]{x^2+1}} = \frac{1}{x^{\frac{2}{3}} \sqrt{1+x^{-2}}} \sim \frac{1}{x^{\frac{2}{3}}}$$

that is $p = \frac{2}{3} < 1$ here. We also have

$$\int_0^\infty \frac{dx}{\sqrt{x^3+1}} < \infty \tag{52}$$

since

$$\frac{1}{\sqrt{x^3+1}} \sim \frac{1}{x^{\frac{3}{2}}}$$

and thus $p = \frac{3}{2} > 1$ in this case. Finally,

$$\int_0^\infty e^{-x^2} dx < \infty$$

because an exponential function (with a negative exponent) decreases (as its argument tends to infinity) faster than any power function (see Sec. IV.14), and therefore comparison test (49) is applicable.

Now we turn to integrals of functions which can assume the values of arbitrary sign:

$$\int_a^\infty f(x)\,dx \tag{53}$$

where either $f(x) \geqslant 0$ or $f(x) \leqslant 0$ for a given x. For such an integral we shall give only one test: if

$$\int_a^\infty |f(x)|\,dx < \infty \tag{54}$$

then integral (53) converges. In this case the integral is said to be **absolutely convergent**, and the function $f(x)$ is called **absolutely integrable** on the interval $a \leqslant x < \infty$.

To prove the test we introduce the functions $f^+(x)$ and $f^-(x)$ which are the "positive and the negative parts" of the function $f(x)$:

$$f^+(x) = \begin{cases} f(x) & \text{for } x \text{ such that } f(x) \geqslant 0 \\ 0 & \text{for } x \text{ such that } f(x) < 0 \end{cases}$$

and

$$f^-(x) = \begin{cases} 0 & \text{for } x \text{ such that } f(x) \geqslant 0 \\ |f(x)| & \text{for } x \text{ such that } f(x) < 0 \end{cases}$$

These functions are shown in Fig. 266. We can write

$$f(x) = f^+(x) - f^-(x) \tag{55}$$

and

$$|f(x)| = f^+(x) + f^-(x)$$

If condition (54) is fulfilled, the area shaded in Fig. 266d is finite. Hence, the areas shaded in Fig. 266b and c are also finite. The inequalities $f^+(x) \geqslant 0$ and $f^-(x) \geqslant 0$ holding, the improper integrals
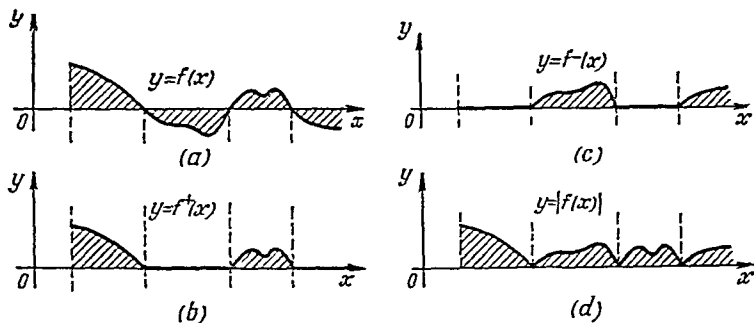
Fig. 266

of these functions converge. Now, by equality (55), integral (53) also converges which is what we set out to prove. More precisely,

$$\int_a^\infty f(x)\,dx = \int_a^\infty f^+(x)\,dx - \int_a^\infty f^-(x)\,dx$$

It can happen t   rges, that is it equals infinity,
          integral   is the so-called conditional
          ce. In   verges but not absolutely.
          case   6b and c are infinite but
          e ti   we take into account
          the   which leads to
          i.

For example, the integral

$$\int\limits_{1}^{\infty} \frac{\sin x}{x^2}\, dx \qquad (56)$$

is absolutely convergent because

$$\left|\frac{\sin x}{x^2}\right| \leqslant \frac{1}{x^2}$$

and therefore the integral of the absolute value of the integrand can be compared with integral (51) for $p = 2$.

If we perform the same operation on the integral

$$\int\limits_{1}^{\infty} \frac{\sin x}{x}\, dx \qquad (57)$$

we shall arrive at integral (51) with $p = 1$ which diverges. Therefore the comparison test does not apply here [see inequality (49) and think why the test is inapplicable]. At the same time it is possible to prove that

$$\int\limits_{1}^{\infty} \left|\frac{\sin x}{x}\right| dx = \int\limits_{1}^{\infty} |\sin x|\, \frac{1}{x}\, dx = \infty$$

because the first factor under the integral sign oscillates about the positive mean value. But it turns out that integral (57) nevertheless converges. To prove this let us perform integration by parts denoting $\frac{1}{x} = u$, $du = -\frac{1}{x^2}\, dx$, and $\sin x\, dx = dv$, $v = -\cos x$:

$$\int\limits_{1}^{\infty} \frac{\sin x}{x}\, dx = -\frac{\cos x}{x}\Big|_{1}^{\infty} - \int\limits_{1}^{\infty} \frac{\cos x}{x^2}\, dx = \cos 1 - \int\limits_{1}^{\infty} \frac{\cos x}{x^2}\, dx$$

The last integral is of the same type as (56) and therefore it converges. Hence, the original integral is also convergent. In other words, integral (57) converges conditionally but not absolutely.

The above results are automatically extended to integrals of complex functions depending on a real argument (see Sec. VIII.6) and to integrals of the form

$$\int\limits_{-\infty}^{b} f(x)\, dx \qquad (58)$$

which are defined by the relation

$$\int\limits_{-\infty}^{b} f(x)\, dx = \lim_{-M \to -\infty} \int\limits_{-M}^{b} f(x)\, dx$$

integral can be carried out in the following way:

$$\int\limits_a^b f(x)\,dx = \int\limits_a^\alpha f(x)\,dx + \int\limits_\alpha^b f(x)\,dx =$$

$$= \lim_{\varepsilon \to +0} \int\limits_{a+\varepsilon}^\alpha f(x)\,dx + \lim_{\varepsilon \to +0} \int\limits_\alpha^{b-\varepsilon} f(x)\,dx =$$

$$= \lim_{\varepsilon \to +0} [F(\alpha) - F(a+\varepsilon)] + \lim_{\varepsilon \to +0} [F(b-\varepsilon) - F(\alpha)] =$$

$$= F(b-0) - F(a+0)$$

Thus. in this case we can use our usual formula (11) for evaluating the definite integral, and if the substitution of the limits of integration for the argument of the function yields finite results the integral is convergent.

We now suppose that integral (61) has a singularity lying inside the interval of integration, for instance, at the point $x = c$. If an antiderivative $F(x)$ is known then we have

$$\int\limits_a^b f(x)\,dx = \int\limits_a^c f(x)\,dx + \int\limits_c^b f(x)\,dx = [F(c-0) - F(a)] +$$

$$+ [F(b) - F(c+0)] = F(b) - F(a) + [F(c-0) - F(c+0)] \quad (63)$$

Consequently, if the antiderivative has no discontinuities, i.e. if $F(c-0) = F(c+0)$, formula (11) can be applied to evaluating the integral. But if the antiderivative has jump discontinuities we must make the necessary corrections as we have done in formula (63). Finally, if the antiderivative has discontinuities of more complicated types inside the interval of integration, in particular, if it approaches infinity at some points, then the integral diverges. The above rules apply to the case of an arbitrary (finite) number of singularities.

Take an example. The integral $\int\limits_{-1}^2 \frac{1}{\sqrt[3]{x}}\,dx$ can be evaluated as follows:

$$\int\limits_{-1}^2 \frac{1}{\sqrt[3]{x}}\,dx = \int\limits_{-1}^2 x^{-\frac{1}{3}}\,dx = \frac{x^{\frac{2}{3}}}{\frac{2}{3}}\Big|_{-1}^2 = \frac{3}{2}(2^{\frac{2}{3}} - 1) = 0.881$$

because in this case the antiderivative is proportional to $x^{\frac{2}{3}}$, i.e. it is continuous, and therefore the integral converges. The last example in Sec. 4 was treated incorrectly because the antiderivative

$-\dfrac{1}{x}$ which we had there approached infinity on the interval of integration at the point $x = 0$ and therefore the integral was divergent.

In evaluating improper integrals we widely use expansions of their integrands into series of different kinds. If such an expansion yields a good approximation only near a singularity then the integral in question is broken into a sum of two integrals so that one of the integrals should be an improper integral taken over a subinterval containing the singularity and the other integral should be an ordinary (proper) integral. Then the first integral is evaluated by means of a series expansion whereas the second one is computed according to the methods of § 3. For example, to evaluate integral (52) we can perform the following operation:

$$\int_0^\infty \frac{dx}{\sqrt{x^3+1}} = \int_0^a \frac{dx}{\sqrt{x^3+1}} + \int_a^\infty x^{-\frac{3}{2}} \left(1+\frac{1}{x^3}\right)^{-\frac{1}{2}} dx =$$

$$= \int_0^a \frac{dx}{\sqrt{x^3+1}} + \int_a^\infty \left(x^{-\frac{3}{2}} - \frac{1}{2} x^{-\frac{9}{2}} + \frac{3}{8} x^{-\frac{15}{2}} - \ldots \right) dx =$$

$$= \int_0^a \frac{dx}{\sqrt{x^3+1}} + \frac{2}{\sqrt{a}} - \frac{1}{7 \sqrt{a^7}} + \frac{3}{52 \sqrt{a^{13}}} - \ldots =$$

$$= \int_0^a \frac{dx}{\sqrt{x^3+1}} + S \qquad (64)$$

Here, in expanding the integrand, we have utilized Newton's second binomial formula (IV.60). The number $a > 0$ entering into (64) can be chosen arbitrarily. If we take very large a we shall encounter difficulties in evaluating the integral $\displaystyle\int_0^a \frac{dx}{\sqrt{x^3+1}}$ but if we take a very small a then the terms of the series whose sum is denoted by S will be very large, and it will be difficult to evaluate the sum. Let us take $a = 2$ and evaluate the last integral by means of Simpson's formula (see Sec. 13). We divide the interval of integration into eight parts. This yields

$$\int_0^2 \frac{dx}{\sqrt{x^3+1}} = 1.402$$

Calculating S with the same accuracy we obtain $S = 1.402$. Hence, integral (52) is equal to 2.804 (let the reader verify all the calculations!).

When we deal with a divergent integral, for instance, of form (41), we can encounter the problem of characterizing the behaviour of its "finite part" (42), which approaches infinity as $N \to \infty$, in a more precise manner. The investigation can also be carried out by means of expansions into series. For example,

$$\int_0^N \frac{dx}{\sqrt[3]{x^2+1}} = \int_0^a \frac{dx}{\sqrt[3]{x^2+1}} + \int_a^N x^{-\frac{2}{3}}(1+x^{-2})^{-\frac{1}{3}} dx =$$

$$= \int_0^a \frac{dx}{\sqrt[3]{x^2+1}} + \int_a^N \left(x^{-\frac{2}{3}} - \frac{1}{3} x^{-\frac{8}{3}} + \frac{2}{9} x^{-\frac{14}{3}} - \dots \right) dx =$$

$$= 3N^{\frac{1}{3}} + C + \frac{1}{5} N^{-\frac{5}{3}} - \frac{2}{33} N^{-\frac{11}{3}} + \dots$$

where $C = \int_0^a \frac{dx}{\sqrt[3]{x^2+1}} - a^{\frac{1}{3}} - \frac{1}{5} a^{-\frac{5}{3}} + \frac{2}{33} a^{-\frac{11}{3}} - \dots$ is a constant which can be calculated by means of the technique applied to the previous example.

An analogous problem can arise when we investigate a convergent integral. In this case integral (41) tends to a finite limit and it is the rate of its variation in the process of approaching the limit that can be investigated by means of expansions into series. For example, reasoning as we did in performing calculations (64), we obtain

$$\int_0^N \frac{dx}{\sqrt{x^3-1}} = \int_0^\infty \frac{dx}{\sqrt{x^3+1}} - \int_N^\infty \frac{dx}{\sqrt{x^3+1}} = 2.804 - \frac{2}{\sqrt{N}} + \frac{1}{7\sqrt{N^7}} - \dots$$

Integrating by parts we find

$$\int_0^N e^{-x^2} dx = \int_0^\infty e^{-x^2} dx - \int_N^\infty \frac{1}{2x} \cdot 2xe^{-x^2} dx =$$

$$= \int_0^\infty e^{-x^2} dx + \frac{1}{2x} e^{-x^2} \Big|_N^\infty + \frac{1}{2} \int_N^\infty \frac{e^{-x^2}}{x^2} dx = \int_0^\infty e^{-x^2} dx -$$

$$- \frac{1}{2N} e^{-N^2} + a \text{ quantity of the order of } \frac{1}{N^3} e^{-N^2}$$

The above estimation can also be easily proved with the help of L'Hospital's rule which we leave to the reader. To specify the expansion we can perform repeated integration by parts.

17. Gamma Function. As an important example of an improper integral let us consider the non-elementary gamma function intro-

duced by Euler in 1729:

$$\Gamma(p) = \int\limits_0^\infty e^{-x} x^{p-1}\, dx \qquad (65)$$

This expression is also called **Euler's integral of the second kind.**
   Integral (65) is improper since it has infinity as its upper limit
of integration and, besides, it has a singularity at $x = 0$ for $p < 1$.
We know that e^{-x} tends to zero, as $x \to \infty$, faster than any negative
power of x and therefore the behaviour of the integrand at $x = \infty$
does not affect the convergence or divergence of the integral. On the
other hand, we have $e^{-x} x^{p-1} \sim \dfrac{1}{x^{1-p}}$ for $x \to 0$ and therefore, by
the beginning of Sec. 16, integral (65) converges for $1 - p < 1$,
i.e. for $p > 0$, and diverges for $1 - p \geqslant 1$, i.e. for $p \leqslant 0$. Therefore
we shall consider formula (65) for $0 < p < \infty$.
   To deduce the basic property

$$\Gamma(p + 1) = p\Gamma(p) \qquad (66)$$

of the gamma function we integrate by parts:

$$\Gamma(p+1) = \int\limits_0^\infty e^{-x} x^{(p+1)-1}\, dx = \int\limits_0^\infty e^{-x} x^p\, dx = -e^{-x} x^p \big|_0^\infty + \int\limits_0^\infty e^{-x} p x^{p-1}\, dx$$

which implies (66).
   Further, we readily find

$$\Gamma(1) = \int\limits_0^\infty e^{-x} x^{1-1}\, dx = \int\limits_0^\infty e^{-x}\, dx = -e^{-x} \big|_0^\infty = 1$$

If we now substitute, in succession, $p = 1, 2, 3, \ldots$ into formula
(66) we obtain

$$\Gamma(2) = 1 \cdot \Gamma(1) = 1, \quad \Gamma(3) = 2\Gamma(2) = 2 \cdot 1,$$

$$\Gamma(4) = 3\Gamma(3) = 3 \cdot 2 \cdot 1 \quad \text{etc.}$$

Generally,

$$\Gamma(n + 1) = n! \quad (n = 1, 2, 3, \ldots) \qquad (67)$$

   Thus we see that the gamma function yields a representation of
the factorial function. At the same time the gamma function also
assumes certain values for non-integral values of the argument and
therefore it extends the factorial function (see Sec. I.15) from discrete
values of the argument to the continuous range of the argument.
The graph of the function is shown in Fig. 268. The equality
$\Gamma(+0) = +\infty$ suggested by formula (66) is also illustrated in
the figure.

In particular, formula (67) implies that

$$0! = \Gamma(1) = 1$$

Further, we have

$$\left(-\frac{1}{2}\right)! = \Gamma\left(-\frac{1}{2} + 1\right) = \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} = 1.772$$

We shall establish the last relation $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ in Sec. 18 [see formula (71)]. It follows that

$$\left(\frac{1}{2}\right)! = \Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2} = 0.886,$$

$$\left(\frac{3}{2}\right)! = \Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3\sqrt{\pi}}{4} = 1.329 \text{ etc.}$$

The gamma function can also be defined for the negative values of the argument but it is impossible to use formula (65) for this purpose since the integral diverges for $p < 0$. However, we can rewrite formula (66) in the form

$$\Gamma(p) = \frac{\Gamma(p+1)}{p} \qquad (68)$$

and use it for defining the gamma function for negative p.
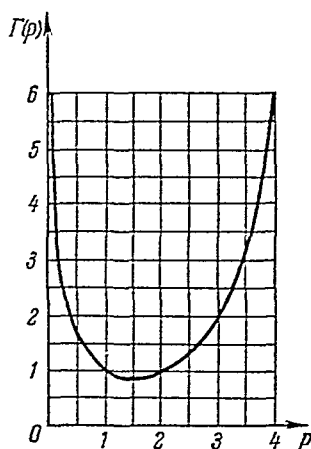
Fig. 268
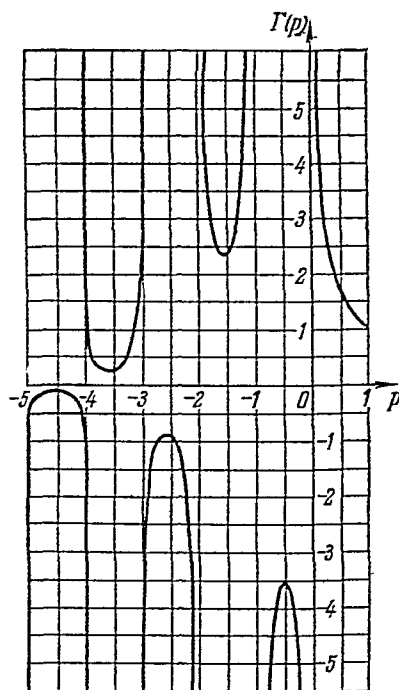
Fig. 269

Indeed, if $-1 < p < 0$ then $0 < p + 1 < 1$ and therefore the right-hand side of (68) makes sense for $-1 < p < 0$. Hence, for-

mula (68) defines the values of the gamma function $\Gamma (p)$ for such p. It should be noted that in this way we obtain $\Gamma (p) < 0$ for $-1 < < p < 0$. Further, if we take $-2 < p < -1$ then $-1 < p + 1 < 0$ and therefore the right-hand side is defined for these p by the preceding extension of the gamma function into the interval $-1 < < p < 0$. By the way, we see that $\Gamma (p) > 0$ for $-2 < p < -1$. We next define $\Gamma (p)$ for $-3 < p < -2$ in the same way etc. Hence, $\Gamma (p)$ has been defined for the values of p of arbitrary signs, and formula (66) holds for all p. Applying formula (68) we conclude, in succession, that $\Gamma (0) = \pm \infty$, $\Gamma (-1) = \pm \infty$, $\Gamma (-2) = \pm \infty$ etc. The graph of the gamma function for the negative values of the argument is shown in Fig. 269.

There are extensive tables of the gamma function. In particular, they can be found in [23].

18. Beta Function. The beta function, or Euler's integral of the first kind, is defined by the formula

$$B (p, q) = \int\limits_0^1 x^{p-1} (1 - x)^{q-1}\, dx \tag{69}$$

Here we must have $p > 0$ and $q > 0$ since otherwise the behaviour of the integrand as x approaches the upper and the lower limits of integration yields the divergence of the integral (why is it so?). It should be noticed that indefinite integral (69) is expressible in terms of elementary functions only for certain specific combinations of the values of the exponents p and q (see Sec. XIII.9).

As we shall show in Sec. XVI.17, the beta function is expressed in terms of the gamma function by the formula

$$B (p, q) = \frac{\Gamma (p)\, \Gamma (q)}{\Gamma (p+q)} \tag{70}$$

The formula suggests an interesting corollary: if we put $p = q = \frac{1}{2}$ we obtain

$$\Gamma^2 \left(\frac{1}{2}\right) = B \left(\frac{1}{2}, \frac{1}{2}\right) = \int\limits_0^1 x^{-\frac{1}{2}} (1 - x)^{-\frac{1}{2}}\, dx =$$

$$= \int\limits_0^1 \frac{dx}{\sqrt{x (1 - x)}} = 2 \int\limits_0^1 \frac{dx}{\sqrt{1 - (1 - 2x)^2}} = - \arcsin (1 - 2x) \,|_0^1 = \pi$$

Hence, the inequality $\Gamma (p) > 0$ holding for $p > 0$, we have

$$\Gamma \left(\frac{1}{2}\right) = \sqrt{\pi} \tag{71}$$

From this we can deduce the value of an important integral of the form $\int_0^\infty e^{-x^2}\,dx$ by using the substitution $x = \sqrt{t}$:

$$\int_0^\infty e^{-x^2}\,dx = \frac{1}{2}\int_0^\infty e^{-t}t^{-\frac{1}{2}}\,dt = \frac{1}{2}\int_0^\infty e^{-t}t^{\frac{1}{2}-1}\,dt = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2} \quad (72)$$

Many definite integrals containing power and exponential functions which cannot be expressed in terms of elementary functions for arbitrary values of parameters involved can often be expressed in terms of the beta function and, hence, in terms of the gamma function. For instance, we have

$$\int_0^{\frac{\pi}{2}} \sin^p x\,dx = \int_0^1 \frac{1}{2}t^{\frac{p}{2}-\frac{1}{2}}(1-t)^{-\frac{1}{2}}\,dt =$$

$$= \frac{1}{2}\mathrm{B}\left(\frac{p+1}{2},\,\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2}\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}+1\right)} \quad \text{for } p > -1$$

(we have used the substitution $\sin x = \sqrt{t}$ here),

$$\int_0^\infty \frac{x^{p-1}}{(1+x)^{p+q}}\,dx = \int_0^1 \frac{y^{p-1}(1-y)^{p+q}}{(1-y)^{p-1}}\frac{dy}{(1-y)^2} =$$

$$= \int_0^1 y^{p-1}(1-y)^{q-1}\,dy = \mathrm{B}(p,q) \quad \text{for } p > 0,\ q > 0 \quad (73)$$

where $x = \frac{y}{1-y}$. Further, we have

$$\int_0^\infty \frac{dx}{(1+x^p)^q} = \frac{1}{p}\int_0^\infty \frac{y^{\frac{1}{p}-1}}{(1+y)^q}\,dy =$$

$$= \frac{1}{p}\mathrm{B}\left(\frac{1}{p},\,q-\frac{1}{p}\right) \quad \text{for } p > 0,\ qp > 1 \quad (74)$$

where $x = y^{\frac{1}{p}}$ [we have utilized formula (73) here]. In particular, (74) yields the value of integral (52):

$$\int_0^\infty \frac{dx}{\sqrt{1+x^3}} = \frac{1}{3}\mathrm{B}\left(\frac{1}{3},\,\frac{1}{2}-\frac{1}{3}\right) = \frac{1}{3}\mathrm{B}\left(\frac{1}{3},\,\frac{1}{6}\right) =$$

$$= \frac{1}{3}\frac{\Gamma\left(\frac{1}{3}\right)\cdot\Gamma\left(\frac{1}{6}\right)}{\Gamma\left(\frac{1}{2}\right)} = \frac{2.676 \times 5.566}{3\sqrt{\pi}} = 2.804$$

(the values of the gamma function have been taken from the tables).

19. Principal Value of Divergent Integral. It is sometimes advisable to attribute certain numerical values to some divergent integrals, in a conditional sense. For instance, such a situation may occur in investigating physical processes in continuous media. There are different ways of employing these values. Let us consider Cauchy's method. Suppose an integral of the form

$$\int_a^b f(x)\,dx \tag{75}$$

has only one singularity inside the interval of integration, say at the point $x = c$. We cut off a subinterval containing the singularity which is symmetric with respect to c. Then we pass to the limit and put, by definition,

$$\text{v.p.} \int_a^b f(x)\,dx = \lim_{\varepsilon \to +0} \left[\int_a^{c-\varepsilon} f(x)\,dx + \int_{c+\varepsilon}^b f(x)\,dx \right] \tag{76}$$

Such a limit can exist even if integral (75) is divergent in the ordinary sense of definition given in Sec. 16. If the limit exists we call (76) the **principal value (Cauchy's principal value)** of (75). The notation v.p. in (76) is the abbreviation of the French *valeur principale* principal value. An integral of this type is often called a singular integral to distinguish it from proper and convergent improper integrals which are called regular integrals. Similarly, the principal value of an improper integral taken over the whole axis of the argument of the integrand is introduced, by definition, as

$$\text{v.p.} \int_{-\infty}^{\infty} f(x)\,dx = \lim_{N \to \infty} \int_{-N}^N f(x)\,dx$$

For instance, the integral $\int_{-1}^2 \dfrac{1}{x}\,dx$ is divergent because the antiderivative $\ln |x|$ of the integrand has a discontinuity at the point $x = 0$ belonging to the interval of integration at which it approaches infinity. At the same time its principal value

$$\text{v.p.} \int_{-1}^2 \frac{1}{x}\,dx = \lim_{\varepsilon \to +0} \left[\int_{-1}^{-\varepsilon} \frac{1}{x}\,dx + \int_{\varepsilon}^2 \frac{1}{x}\,dx \right] =$$

$$= \lim_{\varepsilon \to +0} [\ln |x| \big|_{-1}^{-\varepsilon} + \ln |x| \big|_{\varepsilon}^2] =$$

$$= \lim_{\varepsilon \to +0} [\ln \varepsilon - \ln 1 + \ln 2 - \ln \varepsilon] = \ln 2 = 0.693$$

exists because the summands $\pm \ln \varepsilon$ mutually cancel out before we pass to the limit. Another example is the integral $\int\limits_{-\infty}^{\infty} \sin x\, dx$:

$$\text{v.p.} \int\limits_{-\infty}^{\infty} \sin x\, dx = \lim_{N\to\infty} \int\limits_{-N}^{N} \sin x\, dx =$$

$$= \lim_{N\to\infty} (-\cos x)\big|_{-N}^{N} = \lim_{N\to\infty} [-\cos N + \cos N] = 0$$

Thus, the principal value of the integral exists although the integral diverges, that is it is singular.

It is apparent that not all divergent integrals possess principal values.

§ 5. Integrals Dependent on Parameters

20. Proper Integrals. Take an integral of the form

$$I = \int\limits_{a}^{b} f(x, \lambda)\, dx \tag{77}$$

whose integrand depends on a **parameter** (arbitrary constant) λ besides the variable of integration x. The parameter λ is regarded as constant in the process of integration but generally it can assume different values for which integral (77) is evaluated. And, generally speaking, the result of the integration can also depend on λ, i.e. $I = I(\lambda)$. Such integrals occur in applications when an integrand can involve such parameters as masses, sizes etc. which are kept constant in the process of integration. For the sake of simplicity, we shall take integrands which only depend on one parameter although similar results are obtained in the case of many parameters.

We first consider several formal examples:

$$\int\limits_{0}^{1} (x^2 + \lambda x)\, dx = \frac{1}{3} + \frac{\lambda}{2},$$

$$\int\limits_{0}^{\pi} \sin \alpha x\, dx = \frac{1 - \cos \alpha}{\alpha} \quad \text{and}$$

$$\int\limits_{0}^{1} (s + 1)\, x^s\, dx = 1 \quad (s > -1)$$

In this section we shall take proper integrals of form (77), that is the limits of integration and the integrand will be finite.

Let us consider some properties of these integrals.

1. If the integrand is a continuous function of λ for $a \leqslant x \leqslant b$ then the integral I is also continuous in λ. For example, this is implied by geometric meaning of an integral as the area of a curvilinear trapezoid: if an infinitesimal variation of λ yields an infinitesimal change of the form and of the sizes of the curvilinear side of the trapezoid (which is the graph of the integrand) then the area should also gain an infinitesimal increment.

It should be remarked that at the same time the function f may not be continuous in x and may have finite discontinuities.

We sometimes encounter integrals whose limits of integration can also depend on a parameter:

$$I(\lambda) = \int_{a(\lambda)}^{b(\lambda)} f(x, \lambda)\, dx \tag{78}$$

Then. for $I(\lambda)$ to be continuous, it is sufficient to set the additional condition that the functions $a(\lambda)$ and $b(\lambda)$ have no discontinuities.

2. The *Leibniz formula*

$$\frac{dI}{d\lambda} = \int_a^b f'_\lambda(x, \lambda)\, dx \tag{79}$$

suggests that it is permissible to differentiate integral (77) with respect to the parameter under the integral sign. The matter is that integral (77) is analogous to a sum of a great number of very small summands (see Sec. 2), each of them depending on λ. Since the term-wise differentiation of a sum is permissible because the derivative of a sum equals the sum of the derivatives, formula (78) can be deduced by passing to the limit.

Here we understand formula (79) in the simplest sense, namely, we suppose that integral (77) is proper and integral (79) is proper or improper but convergent. But there are cases when integral (79) diverges. Then formula (79) remains true provided we understand it in a generalized sense which will be discussed in Sec. 27.

When differentiating integral (78) we must take into account that λ enters into the formula as a parameter on which the integrand depends and as a variable on which both upper and lower limits of integration depend. Therefore we must use the formula for the derivative of a composite function (see Sec. IX.12) and formulas for the derivatives of an integral with respect to its upper and lower limits (see Sec. 4). This implies

$$\frac{dI}{d\lambda} = \int_{a(\lambda)}^{b(\lambda)} f'_\lambda(x, \lambda)\, dx + f(b(\lambda), \lambda)\, b'(\lambda) - f(a(\lambda), \lambda)\, a'(\lambda) \tag{80}$$

3. When we integrate an integral dependent on a parameter with respect to the parameter, it is permissible to integrate the integrand under the sign of the integral in (77), that is

$$\int\limits_{\alpha}^{\beta} I(\lambda) \, d\lambda = \int\limits_{a}^{b} \left(\int\limits_{\alpha}^{\beta} f(x, \lambda) \, d\lambda \right) dx$$

The assertion is justified in the same way as property 2.

21. Improper Integrals. We shall take the case of integrals of the form

$$I(\lambda) = \int\limits_{a}^{\infty} f(x, \lambda) \, dx \tag{81}$$

which have no singularities for finite x. Improper integrals of other types (see Sec. 16) dependent on parameters possess similar properties. Obviously, we suppose that integral (81) converges. But contrary to Sec. 20, here we can have a situation when the dependence of an integral on a parameter λ may not be continuous even if the function f is continuous in λ which is a new fact for us. This is possible because even an infinitesimal variation of a function over an infinite interval of integration may lead to a finite variation of the value of the integral.

For instance, we shall show in Sec. XVII.32 that

$$\int\limits_{0}^{\infty} \frac{\sin x}{x} \, dx = \frac{\pi}{2}$$

It immediately follows that for $\lambda > 0$ we have

$$I(\lambda) = \int\limits_{0}^{\infty} \frac{\sin \lambda x}{x} \, dx = \int\limits_{0}^{\infty} \frac{\sin s}{s} \, ds = \frac{\pi}{2}$$

(we have made the substitution $\lambda x = s$). At the same time we have $I = 0$ for $\lambda = 0$ and, for $\lambda < 0$, we obtain

$$I(\lambda) = \int\limits_{0}^{\infty} \frac{\sin(-|\lambda|) x}{x} \, dx = - \int\limits_{0}^{\infty} \frac{\sin |\lambda| x}{x} \, dx = - \frac{\pi}{2}$$

Hence, in this example $I(\lambda)$ has a jump discontinuity at $\lambda = 0$. One can find it strange that $I(\lambda)$ is discontinuous because the value of the improper integral

$$I(\lambda) = \lim_{N \to \infty} \int\limits_{0}^{N} \frac{\sin \lambda x}{x} \, dx$$

is obtained as the limit of the values of the corresponding proper integrals, and each of these proper integrals is continuous in λ. The explanation is that the limit of a sequence of continuous functions may not be a continuous function, as it will be shown here.

We now take the functions

$$I_N(\lambda) = \int_0^N \frac{\sin \lambda x}{x}\, dx$$

whose graphs are depicted in Fig. 270 for small values of N and for large values of N. These functions are continuous in λ but the
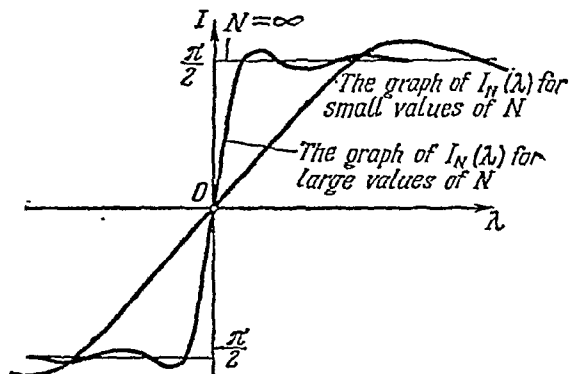
Fig. 270

transition from the values which are close to $-\frac{\pi}{2}$ to the values which are close to $\frac{\pi}{2}$ takes place on a small interval of variation of λ for large N. The greater N, the smaller the interval. Therefore in the limiting process, for $N = \infty$, this transition takes place on an infinitesimal interval of the λ-axis, that is there appears a discontinuity.

The possibility of the existence of such discontinuities complicates the investigation of integrals of form (81) and, particularly, it makes it impossible to apply directly properties 2 and 3 in Sec. 20. Therefore we sometimes have to replace integral (81) by an integral taken over a finite interval from a to N and then pass to the limit as $N \to \infty$. Nevertheless, as we shall show, there exists an important particular case when such discontinuities are impossible.

Suppose that the function $f(x, \lambda)$ satisfies the conditions

$$|f(x, \lambda)| \leqslant \varphi(x) \quad (a \leqslant x < \infty) \quad \text{and} \quad \int_a^\infty \varphi(x)\, dx < \infty \quad (82)$$

for all the values of λ in question where $\varphi(x)$ is a certain function. Then, by Sec. 15, integral (81) converges for all λ. In this case we shall say that the convergence of the integral is regular. For instance, this is the case for the integral

$$\int_1^\infty \frac{\sin \lambda x}{x^2}\, dx$$

since

$$\left|\frac{\sin \lambda x}{x^2}\right| \leqslant \frac{1}{x^2} \quad \text{and} \quad \int_1^\infty \frac{1}{x^2}\, dx = 1 < \infty$$

We can assert that if the integrand of integral (81) depends continuously on λ in the case of regular convergence the integral I is continuous in λ. In fact, the integral can be represented in the form

$$I(\lambda) = \int_a^N f(x, \lambda)\, dx + \int_N^\infty f(x, \lambda)\, dx$$

The first summand is a proper integral and it is therefore continuous in λ. The second summand can be estimated as

$$\left|\int_N^\infty f(x, \lambda)\, dx\right| \leqslant \int_N^\infty |f(x, \lambda)|\, dx \leqslant \int_N^\infty \varphi(x)\, dx \quad .$$

and, by condition (82), this integral becomes small for all the values of λ simultaneously, for sufficiently large N (see Sec. 15). The variation of the whole sum $I(\lambda)$ corresponding to a small variation of λ must therefore be small which means that I depends continuously on λ.

The properties of regularly convergent integrals are completely analogous to those of proper integrals described in Sec. 20.

§ 6. Line Integrals

22. Line Integrals of the First Type. The third example in Sec. 1 is an example of a line integral of this type. The general definition is formulated as follows.

Suppose there is a curve (L) of finite length lying in space or in a plane. Let a quantity u be determined at each point of the curve. If we reckon the arc length s along the curve (L) from a certain point of the curve then we can regard u as a function of s: $u = f(s)$. To form an integral sum we break up the curve (L) into small elementary arcs. Let the points of division correspond to the values

$$\alpha = s_0 < s_1 < \ldots < s_n = \beta$$

DEFINITE INTEGRAL

where $s = \alpha$ and $s = \beta$ are, respectively, the values corresponding to the ends of the curve (L), and n is the number of elementary arcs $s_{k-1} \leqslant s \leqslant s_k$. Now we choose an arbitrary point σ_k on each elementary arc (that is $s_{k-1} \leqslant \sigma_k \leqslant s_k$) and form the integral sum

$$\sum_{k=1}^{n} f(\sigma_k) \Delta s_k, \quad \Delta s_k = s_k - s_{k-1}$$

To obtain the integral we must pass to the limit in the process when all the lengths of the elementary arcs are decreased unlimitedly (compare with Sec. 2), i.e.

$$\int\limits_{(L)} u \, ds = \int\limits_{(L)} f(s) \, ds = \int\limits_{\alpha}^{\beta} f(s) \, ds = \lim \sum_{k=1}^{n} f(\sigma_k) \Delta s_k \qquad (83)$$

It is integral (83) in which the integration is performed with respect to the arc length that is called a **line integral of the first type**. For instance, in the above-mentioned example from Sec. 1 we have

$$M = \int\limits_{(L)} \rho \, ds$$

Similarly, if a point describes a trajectory (L) and a force $\mathbf{F}$ is acting on the point in this motion ($\mathbf{F}$ can be variable in the general case) then the work performed by the force is (compare with the corresponding example in Sec. 6)

$$A = \int\limits_{(L)} F \cos(\widehat{\mathbf{F}, \tau}) \, ds$$

where τ is the unit vector in the direction of the tangent to (L). Hence, this work is a line integral of the function $f = F \cos(\widehat{\mathbf{F}, \tau})$ with respect to the arc length.

Formula (83) indicates that a line integral of the first type is a modification of the ordinary definite integral, and therefore many properties of the definite integral (in particular, properties 2-5 in Sec. 4 and those enumerated in Sec. 5) are automatically extended to the line integral. But at the same time it should be taken into account that Δs, and therefore ds, is always considered to be positive which implies that when we pass from (83) to an ordinary definite integral we must perform integration from a smaller value of s to a larger value. Therefore property 1 in Sec. 4 makes no sense in the case of integrals with respect to the arc length.

A quantity u can sometimes be defined over the whole space. For instance, it may be represented by a function $u = f(x, y, z)$. Then integral (83) can be put down in the form

$$I = \int\limits_{(L)} f(x, y, z) \, ds$$

If a curve (L) is represented parametrically in the form $x = x(t)$, $y = y(t)$ and $z = z(t)$ the evaluation of the corresponding integral can be performed according to the formula

$$I = \int\limits_{\gamma}^{\delta} f(x(t),\ y(t),\ z(t)) \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\ dt$$

where the values $t = \gamma$ and $t = \delta$ correspond to the ends of the curve (L). Here we have taken the expression of ds given in Sec. VII.23. On the basis of Sec. VII.23, we can also write an integral of the first type in the form $\int\limits_{(L)} u\ |dr|$. We also write it as

$\int\limits_{(L)} u(M)\ ds$ where M is the variable point of the curve (L). The corresponding integral sum can also be written in the form

$$\sum_{k=1}^{n} u_k \Delta s_k = \sum_{k=1}^{n} u(M_k)\ \Delta s_k$$

where M_k is a point belonging to the kth elementary arc.

Let us take an example illustrating an application of the line integral of the first type. It is well known in mechanics that if we are given a system of material points $M_k(x_k, y_k)$ of masses m_k lying in the x, y-plane where $k = 1, 2, \ldots, n$, then the coordinates of the centre of gravity of the system are defined by the formulas

$$x_c = \frac{m_1 x_1 + m_2 x_2 + \ldots + m_n x_n}{m_1 + m_2 + \ldots + m_n}, \qquad y_c = \frac{m_1 y_1 + m_2 y_2 + \ldots + m_n y_n}{m_1 + m_2 + \ldots + m_n}$$

Now suppose that we have a plane material line (L) with linear density ρ which may be variable in the general case. The centre of gravity of the material curve can be found in the following way.

Let us divide (mentally) the curve (L) into small elementary arcs Δs_k and replace each of the arcs by the material point of mass $m_k = \rho_k \Delta s_k$ lying on the arc. Thus we obtain a "discrete model" of the material line. The centre of gravity of the model has the coordinates

$$x_c = \frac{\sum\limits_{k=1}^{n} x_k \rho_k \Delta s_k}{\sum\limits_{k=1}^{n} \rho_k \Delta s_k}, \qquad y_c = \frac{\sum\limits_{k=1}^{n} y_k \rho_k \Delta s_k}{\sum\limits_{k=1}^{n} \rho_k \Delta s_k} \tag{84}$$

Now if we pass to the limit, as the lengths of the elementary arcs are decreased unlimitedly, our model will turn into the continuous curve (L) whose centre of gravity, by formulas (84), will have the

coordinates

$$x_c = \frac{\int\limits_{(L)} \rho x \, ds}{\int\limits_{(L)} \rho \, ds}, \qquad y_c = \frac{\int\limits_{(L)} \rho y \, ds}{\int\limits_{(L)} \rho \, ds} \qquad (85)$$

If we take a curve with constant linear density the formulas for the centre of gravity are simplified. Namely, cancelling out $\rho = $ = const in formulas (85) we obtain the formulas for the so-called geometric centre of gravity of the curve (L):

$$x_{g.c.} = \frac{\int\limits_{(L)} x \, ds}{L}, \qquad y_{g.c.} = \frac{\int\limits_{(L)} y \, ds}{L} \qquad (86)$$

where L designates the length of the curve (L).

Let us compare the second formula (86) with formula (36) of the area of the surface of a solid of revolution. We see that

$$S = 2\pi \int\limits_{(L)} y \, ds = L \cdot 2\pi y_{g.c.}$$

In other words, if a plane curve rotates about an axis lying in the plane of the curve and not intersecting it then the area of the surface
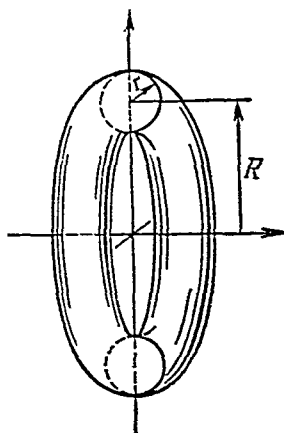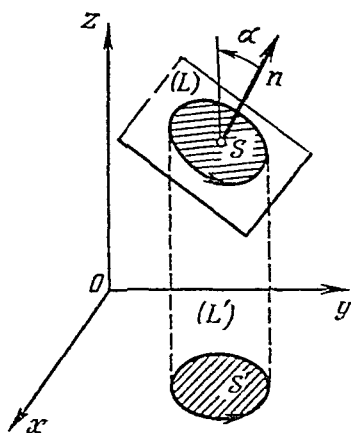
Fig. 271   Fig. 272

of revolution thus obtained is equal to the product of the length of the curve by the distance passed by its geometric centre of gravity. This is Guldin's first theorem named after the Swiss mathematician P. Guldin (1577-1643) who applied the theorem. For instance, the theorem implies that the surface area of a torus (see Fig. 271), obtained by the rotation of a circle of radius r about an axis lying in the

31—0141

same plane as the circle and not intersecting it, is equal to

$$S = 2\pi r \cdot 2\pi R = 4\pi^2 rR$$

where R is the distance from the centre of the circle to the axis of revolution.

We shall come back to the notion of a line integral of the first type in Sec. XVI.1.

23. Line Integrals of the Second Type. There are line integrals taken with respect to coordinates, besides the integrals of the first type. When forming such an integral (called a **line integral of the second type**) we suppose that the curve (L) is directed, that is there is an indication in what direction the curve is described. If we have a non-closed curve we must indicate which of its ends is regarded as its initial point and which as its terminal point. To define the integrals we write, instead of formula (83), the formulas

$$\left.\begin{aligned}
\int\limits_{(L)} u\, dx &= \lim \sum_{h=1}^{n} f(\sigma_h)\, \Delta x_h, \\
\int\limits_{(L)} u\, dy &= \lim \sum_{k=1}^{n} f(\sigma_k)\, \Delta y_k \text{ and } \int\limits_{(L)} u\, dz = \lim \sum_{h=1}^{n} f(\sigma_k)\, \Delta z_k
\end{aligned}\right\} \quad (87)$$

where Δx_h is the increment of the abscissa x along the kth elementary arc etc.

Integrals of type (87) are readily reduced to ordinary definite integrals. For example, if the curve (L) is represented in a parametric form then the values of u assumed at the points of (L) become dependent on t, i.e. u becomes a function of t. Therefore

$$\int\limits_{(L)} u\, dx = \int\limits_{\gamma}^{\delta} u(t)\, \dot{x}(t)\, dt$$

where the values $t = \gamma$ and $t = \delta$ correspond to the ends of the curve (L). Consequently, the basic properties of the definite integral are extended to line integrals of the second type (properties 2-5 in Sec. 4). But here property 1 in Sec. 4 also holds. It can be formulated as follows: if the direction of describing the curve (L) is reversed then integrals of type (87) are multiplied by -1. In fact, if the curve (L) is passed in the opposite direction then all Δx (and all dx) change their signs. The possibility of changing the sign of dx also indicates that the properties which are connected with integration of inequalities (see Sec. 5) no longer hold for the integrals of the second type. For instance, an integral of the form (87) of a positive function may not be positive, in contrast to the integral of the first type.

In the theory of differential equations (Sec. XV.6) and in the theory of vector field (§ XVI.6) we often use the combinations of integrals of type (87) of the form

$$\int\limits_{(L)} (u\,dx + v\,dy + w\,dz) = \int\limits_{(L)} u\,dx + \int\limits_{(L)} v\,dy + \int\limits_{(L)} w\,dz$$

where u, v and $\dot{w}$ are given functions of x, y and z.

Some examples considered in Sec. 8 were virtually integrals of type (87). Indeed, formulas (29)-(31) can be rewritten as

$$S = -\int\limits_{(L)} y\,dx = \int\limits_{(L)} x\,dy = \frac{1}{2} \int\limits_{(L)} (x\,dy - y\,dx) \tag{88}$$

on the basis of formulas $\dot{x}\,dt = dx$ and $\dot{y}\,dt = dy$.

For our further aims we shall need the integral $\int\limits_{(L)} y\,dx$ taken along a closed plane curve arbitrarily placed in space, as in Fig. 272. To evaluate the integral we project the curve (L) on the xOy-plane and verify that

$$\int\limits_{(L)} y\,dx = \int\limits_{(L')} y\,dx \tag{89}$$

Indeed, the points of the curves (L) and (L') which correspond to each other differ only in the values of the coordinate z which does not affect integrals (89).

By (88), the integral on the right-hand side of (89) is equal to

$$\int\limits_{(L')} y\,dx = -S'$$

Now we shall use a well-known property of projections: if a plane figure is projected on another plane then the area of the projection is equal to the product of the area of the initial figure by the cosine of the angle between the planes (see

Fig. 273

Fig. 273). Actually, in this case the sizes in one direction are multiplied by $\cos \alpha$ whereas the sizes in the perpendicular direction do not change.

Thus, we have

$$\int\limits_{(L)} y\,dx = -S \cos \alpha = -S \cos (\widehat{\mathbf{n},\,z}) \tag{90}$$
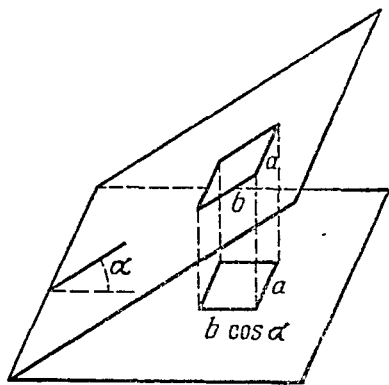
where S is the area bounded by the curve (L) and $\mathbf{n}$ is the unit vector in the direction of the outer normal to the plane of the curve (L). The direction of the vector $\mathbf{n}$ corresponds to the rule of describing (L) according to the right-hand screw rule (see Sec. VII.11).

The formula

$$\int\limits_{(L)} x\, dy = S \cos(\widehat{\mathbf{n}, x}) \tag{91}$$

is proved in a similar way.

Performing circular permutation of the coordinate axes (see Sec. VII.12) we deduce, from formulas (90) and (91), the formulas

$$\left.\begin{array}{l} \int\limits_{(L)} z\, dy = -S \cos(\widehat{\mathbf{n}, x}), \quad \int\limits_{(L)} x\, dz = -S \cos(\widehat{\mathbf{n}, y}), \\[2mm] \int\limits_{(L)} y\, dz = S \cos(\widehat{\mathbf{n}, x}) \quad \text{and} \quad \int\limits_{(L)} z\, dx = S \cos(\widehat{\mathbf{n}, y}) \end{array}\right\} \tag{92}$$

We note in conclusion that integrals of the forms

$$\oint\limits_{(L)} f(x)\, dx, \quad \oint\limits_{(L)} \varphi(y)\, dy \quad \text{and} \quad \oint\limits_{(L)} \psi(z)\, dz$$

taken along any closed curve (L) are always equal to zero (the notation $\oint$ is used for designating integrals taken along closed contours). Actually, let $F(x)$ be an antiderivative of $f(x)$. Then the first of the above integrals is equal to the increment of the function $F(x)$ gained as the variable point describes (L) and returns to its original position. The integral is therefore equal to zero. The second and the third integrals are treated similarly.

24. Conditions for a Line Integral of the Second Type to Be Independent of the Path of Integration. Consider an integral of the form

$$I = \int\limits_{(L)} [P(x, y, z)\, dx + Q(x, y, z)\, dy + R(x, y, z)\, dz] \tag{93}$$

where P, Q and R are some functions defined over the whole space with the coordinates x, y and z or in a domain lying in the space. Let (L) be an arbitrary curve lying inside the domain. We shall also suppose that P, Q and R are finite in the domain, i.e. they are bounded and do not approach infinity at any point. There are such physical problems in which we encounter integrals (93) that only depend on the positions of the initial and terminal points of the curve (L) but not on the form of (L). This means that these integrals are independent of the way in which (L) connects the initial point with the terminal point. In other words, in these cases we have

(Fig. 274)

$$\int\limits_{(L')} (P\,dx + Q\,dy + R\,dz) = \int\limits_{(L'')} (P\,dx + Q\,dy + R\,dz) =$$

$$= \int\limits_{(L''')} (P\,dx + Q\,dy + R\,dz) = \ldots \qquad (94)$$

for any fixed points A and B and for all the possible curves L', L'', L''', ... connecting A and B (where L', L'', L''', ... lie inside the domain). We shall say that integral (93) is *independent of the path of integration*. Such an integral can express, for instance,
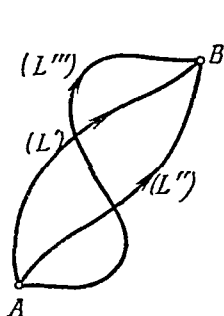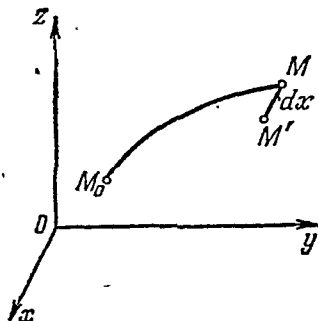
Fig. 274   Fig. 275

the work of a field of force as a point is moving. Then condition (94) means that the work only depends on the positions of the initial and terminal points.

For integral (93) to be independent of the path of integration, it is necessary and sufficient that integral (93) taken round any closed contour should be equal to zero, that is

$$\oint\limits_{(L)} (P\,dx + Q\,dy + R\,dz) = 0 \qquad (95)$$

for any closed contour (L).

To prove the assertion we shall first suppose that condition (95) is fulfilled and that we are given paths (L') and (L'') with the same initial and terminal points (see Fig. 274). Let us construct the closed contour (L) traced from A to B along (L') and from B to A along (L''), the path (L'') being described in the direction which is opposite to its original direction in the corresponding integral (94). Now we take condition (95), break up integral (95) into two integrals taken along the two paths, according to property 3 in Sec. 4, and change the direction of integration for the second integral with the simultaneous change of its sign, according to property 1 in Sec. 4. This

yields

$$0 = \oint\limits_{(L)} = \int\limits_{(L')} - \int\limits_{(L'')}, \quad \text{i.e.} \quad \int\limits_{(L')} (P\,dx + Q\,dy + R\,dz) =$$

$$= \int\limits_{(L'')} (P\,dx + Q\,dy + R\,dz)$$

which implies (94). Reversing the preceding argument we can easily deduce (95) from (94).

. For integral (93) to be independent of the path of integration, it is necessary and sufficient that the element of integration should be the total differential of some single-valued function of three variables. that is

$$P\,dx + Q\,dy + R\,dz \equiv du \tag{96}$$

where $u = u\,(x, y, z)$ is a certain function. To prove the assertion we first suppose that condition (96) holds. Then

$$\int\limits_{(L)} (P\,dx + Q\,dy + R\,dz) = \int\limits_{(L)} du = u\,(B) - u\,(A)$$

where A and B are, respectively. the initial and the terminal points of the curve (L). Consequently, the integral is independent of the path of integration.

Conversely. let integral (93) be independent of the path of integration. We fix an arbitrary point M_0 in space and introduce the following function of the variable point $M\,(x, y, z)$:

$$u\,(M) = \int\limits_{\cup M_0 M} (P\,dx + Q\,dy + R\,dz) \tag{97}$$

where $\cup M_0 M$ is an arbitrary curve connecting M_0 and M and directed from M_0 to M. Condition (94) suggests that the function is single-valued. that is it assumes a certain uniquely defined value at each point M. To find du we first give x an infinitesimal increment dx. Then the point M passes to the position M'. as in Fig. 275, and the infinitesimal line segment MM' is parallel to the x-axis. The corresponding increment of the function u is equal to

$$\Delta_x u = u\,(M') - u\,(M) = \int\limits_{\cup M_0 M M'} (P\,dx + Q\,dy + R\,dz) -$$

$$- \int\limits_{\cup M_0 M} (P\,dx + Q\,dy + R\,dz) = \int\limits_{M M'} (P\,dx + Q\,dy + R\,dz)$$

But the coordinates y and z not varying along the segment MM', we have $dy = dz = 0$ in the last integral. Therefore $\Delta_x u = \int\limits_{MM'} P\,dx$.

The segment MM' being infinitesimal, we can consider P to be constant on it $(P = \text{const})$ to within infinitesimals of higher order of smallness. Therefore, passing from the increment to the differential and thus dropping infinitesimals of higher order we obtain $\partial_x u = P\,dx$. Similarly, we find that $\partial_y u = Q\,dy$ and $\partial_z u = R\,dz$. Adding up the results we get (see Sec. IX.11)

$$du = P\,dx + Q\,dy + R\,dz$$

and consequently condition (96) is fulfilled.

If we recall expression (IX.7) of the total differential we can write condition (96) in the equivalent form

$$\frac{\partial u}{\partial x} = P, \quad \frac{\partial u}{\partial y} = Q, \quad \frac{\partial u}{\partial z} = R \tag{98}$$

It follows that if integral (93) is independent of the path of integration then we have

$$\frac{\partial P}{\partial y} \equiv \frac{\partial Q}{\partial x}, \quad \frac{\partial P}{\partial z} \equiv \frac{\partial R}{\partial x}, \quad \frac{\partial Q}{\partial z} \equiv \frac{\partial R}{\partial y} \tag{99}$$

Indeed, conditions (98) imply

$$\frac{\partial P}{\partial y} \equiv \frac{\partial^2 u}{\partial x\,\partial y}, \quad \frac{\partial Q}{\partial x} \equiv \frac{\partial^2 u}{\partial y\,\partial x}$$

and therefore, based on the independence of the mixed derivatives of the order of differentiation (see Sec. IX.15), we obtain the first equality (99). The other conditions (99) are proved similarly (let the reader complete the proof!).

In the theory of vector field (Sec. XVI.27) we shall prove the converse assertion: if conditions (99) are fulfilled and if the domain in which the functions P, Q and R are considered is simply-connected then integral (93) is independent of the path of integration. A domain is said to be *simply-connected* if any closed contour lying in it can be contracted to a point by means of a continuous deformation without falling outside the domain.

The whole space, a half-space, a dihedral or a polyhedral angle, the interior or the exterior of a sphere, the interior of a finite or infinite circular cylinder are examples of simply-connected domains. In contrast to it, the exterior of an infinite circular cylinder is a *doubly-connected* domain. Indeed, if we consider the contour (L) depicted in Fig. 276 we see that it cannot be contracted to a point without falling outside the domain. Further examples of multiply-connected domains are the interior or the exterior of a torus and the whole space from which all the points belonging to a circle or to an infinite straight line are removed. In Fig. 277 we see a plane simply-connected domain and a plane four-connected domain.

Comparing conditions (96) and (99) we see that conditions (99) are necessary and sufficient for the expression $P\,dx + Q\,dy + R\,dz$ to be a total differential of a single-valued function $u\,(x,\,y,\,z)$ defined in a simply-connected domain. It is possible to show that if conditions (99) are fulfilled in a multiply-connected domain then
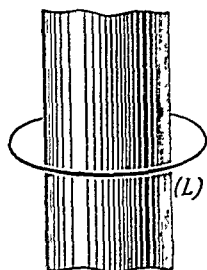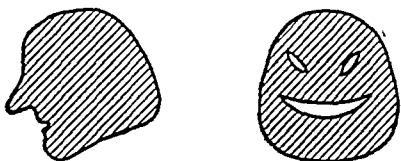
Fig. 276   Fig. 277

the function $u\,(x,\,y,\,z)$ constructed in accordance with formula (97) satisfies condition (96) but may not be single-valued in the general case.

An integral of the form

$$\int [P\,(x,\,y)\,dx + Q\,(x,\,y)\,dy]$$

taken along a plane curve is considered similarly. Accordingly, in all the expressions we must drop the terms containing z and dz. In particular, conditions (99) turn into one condition

$$\frac{\partial P}{\partial y} \equiv \frac{\partial Q}{\partial x}$$

for such an integral.

§ 7. The Concept of Generalized Function

25. Delta Function. The *Dirac delta function* which is widely used in mathematics and its applications is the simplest example of a generalized function. The theory of generalized functions was founded in 1936 by the Soviet mathematician S. L. Sobolev. It was thoroughly developed only during the last twenty years.

To get an approximate representation of the delta function we first consider the discontinuous function defined by the equalities

$$y = \delta_N\,(x) = \begin{cases} 0 \text{ for } -\infty < x < -\frac{1}{2N} \\ N \text{ for } -\frac{1}{2N} < x < \frac{1}{2N} \\ 0 \text{ for } \frac{1}{2N} < x < \infty \end{cases} \tag{100}$$

where N is a very large positive number. The graph of the function is shown in Fig. 278. As usual, the values of the function at the points of discontinuity $x = \pm \frac{1}{2N}$ are of no importance and therefore they are not indicated here. The delta function can be thought of as the limit of the function (100) as $N \to \infty$. But, strictly speaking, such a limit should be defined as

$$\delta(x) = \begin{cases} 0 & \text{for } -\infty < x < -0 \text{ and } +0 < x < \infty \\ \infty & \text{for } -0 < x < +0 \end{cases}$$

with the additional condition

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1$$

which is implied by the fact that $\int_{-\infty}^{\infty} \delta_N(x)\, dx = 1$ because the area shaded in Fig. 278 is equal to unity. We cannot therefore speak about the graph of a delta function. The last relation can also be rewritten in the form

$$\int_{-0}^{+0} \delta x\,(dx) = 1 \qquad (101)$$

Fig. 278

An approximate representation of the delta function must not necessarily be obtained by means of the discontinuous function (100). For example, we can also take the function $\delta_N(x) = \frac{N}{\pi}(1 + N^2 x^2)^{-2}$ defined over the interval $-\infty < x < \infty$ for this purpose (let the reader investigate the behaviour of the graph of this function as $N \to \infty$) and so on. Generally, we can take every function whose values are "concentrated" near $x = 0$. More precisely, for the delta function to be defined as the limit of the function $\delta_N(x)$ (understood in the above sense) it is sufficient that $\delta_N(x)$ should satisfy the conditions $\delta_N(x) \geqslant 0$ for $-\infty < x < \infty$, $\int_a^b \delta_N(x)\, dx \to 0$,

$\int_{-a}^{-b} \delta_N(x)\, dx \to 0$ and $\int_{-a}^{b} \delta_N(x)\, dx \to 1$, as $N \to \infty$ for any positive constant numbers a and b.
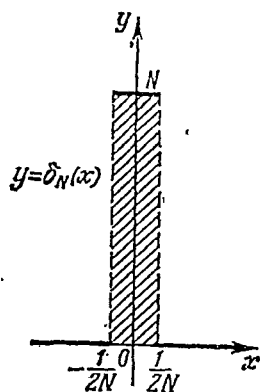
If we consider masses distributed along the x-axis and their linear densities then the linear density of a material point of unit mass placed at the coordinate origin turns out to be equal to the delta function. In fact, if we first consider this mass to be uniformly distributed over the segment $-\frac{1}{2N} \leqslant x \leqslant \frac{1}{2N}$ but not concentrated at the point $x = 0$ then its density will be of the form depicted in Fig. 278 (why is it so?). Now if $N \to \infty$ then the mass will be contracted to the point in the limiting process and its density will become the delta function.

Similarly, the function $\rho(x) = m\delta(x - a)$ is the linear density of a mass m concentrated at the point $x = a$ (we say "linear density" because we regard the mass as distributed along a line, that is along the x-axis in our case). The density of a point charge or of a point force can be represented in a similar way and so on.

It is sometimes necessary to add together a delta function and an ordinary function. For example, the sum

$$\rho(x) = \rho_0 + m_1\delta(x - a_1) + m_2\delta(x - a_2)$$

is the density of the combination of a uniformly distributed mass and two discrete mass points. Hence, the use of the delta function makes it possible to apply formulas which were originally deduced for continuously distributed masses to any combinations of discrete mass points and distributed masses. Moreover, from this point of view the distinction between discrete mass points and continuously distributed masses becomes inessential. The same refers to charges, forces and so on.

When integrating an expression containing delta functions one must apply formula (101). For instance, if $f(x)$ is a continuous function in the interval $\alpha < x < \beta$ and if $\alpha < a < \beta$ then

$$\int_{\alpha}^{\beta} f(x)\,\delta(x-a)\,dx = \int_{\alpha}^{a-0} f(x)\,\delta(x-a)\,dx + \int_{a-0}^{a+0} f(x)\,\delta(x-a)\,dx +$$

$$+ \int_{a+0}^{\beta} f(x)\,\delta(x-a)\,dx = 0 + \int_{a-0}^{a+0} f(a)\,\delta(x-a)\,dx + 0 =$$

$$= f(a)\cdot 1 = f(a) \qquad (102)$$

because the first and the third integrals on the right-hand side are equal to zero since the delta function is equal to zero on the corresponding intervals whereas $f(a)$ substitutes for $f(x)$ in the second integral since a continuous function can be regarded as being constant on an infinitesimal interval. Therefore formula (101) yields (102).

We must indicate the two-fold sense of an integral of the form $\int_{\alpha}^{\beta} f(x) \delta(x - \alpha)\, dx$. We can speak of the value of the integral only if there is a specification as to whether the singularity of the delta function $\delta(x - \alpha)$ (i.e. the point $x = \alpha$) is included into the range of integration or not. Consequently, we must write either

$$\int_{\alpha-0}^{\beta} f(x)\, \delta(x - \alpha)\, dx = f(\alpha) \quad \text{or} \quad \int_{\alpha+0}^{\beta} f(x)\, \delta(x - \alpha)\, dx = 0.$$

Integrating the Dirac delta function from $-\infty$ to x we obtain the step function (the Heaviside unit function)

$$e(x) = \int_{-\infty}^{x} \delta(t)\, dt = \begin{cases} 0 \text{ for } -\infty < x < 0 \\ 1 \text{ for } 0 < x < \infty \end{cases} \tag{103}$$

which also has many applications. Its graph is shown in Fig. 279. Such a function can be applied to describing a process of an instantaneous application of a constant action, for instance, the process which occurs when a constant voltage is instantaneously applied to the terminals of an electric circuit.

Thus, the integration of the delta function results in an ordinary though discontinuous function. The repeated integration yields a continuous function (let the reader investigate the form of the function).

If we differentiate equality (103) we obtain

$$\delta(x) = e'(x) \tag{104}$$

This equality should be understood in the generalized sense. For instance, we can first substitute the oblique segment connecting the points $\left(-\frac{1}{2N},\ 0\right)$ and $\left(\frac{1}{2N},\ 1\right)$ for the vertical segment in Fig. 279 (this inclined segment is represented by the dotted line). Then the discontinuous function is replaced by a continuous function whose derivative has the graph of the form shown in Fig. 278. Now passing to the limit, as $N \to \infty$, we obtain relation (104).

Hence, we can say that a delta function is obtained by differentiating a discontinuous function having a finite jump. For example, let us take the law of motion considered in Sec. I.13 which is connected with Fig. 6. The corresponding velocity is expressed by the formula

$$s'_t = \begin{cases} gt \text{ for } 0 \leqslant t < t^* \\ v \text{ for } t^* < t < \infty \end{cases}$$

and it has the jump $s'_t(t^* + 0) - s'_t(t^* - 0) = gt^* - v$ at the point $t = t^*$. The acceleration is therefore equal to $(gt^* - v)\,\delta\,(t - -t^*) + ge\,(t^* - t)$ (check it up!). The first term in the sum describes the impact phenomenon.

26. Application to Constructing Influence Function. Constructing an influence function (which is also called Green's function after the English mathematician G. Green, 1793-1841) is one of the important applications of the delta function. We begin with an example. Let us consider the deflection $h\,(x)$ of a beam subjected to a transversal external load of intensity $p\,(x)$ (see Fig. 280 where we see
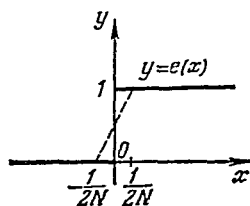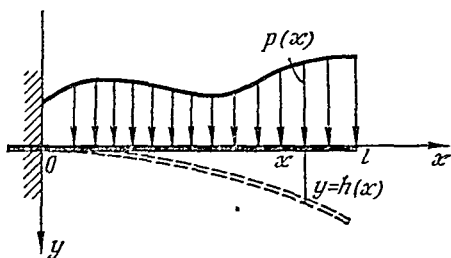
Fig. 279   Fig. 280

the graph or, as it is called, the *diagram of the external load*). We shall suppose that the load is not very large and we can therefore apply the linear law of elasticity which implies that if we combine external forces the corresponding deflections add up.

Let us imagine that we have unit force applied at a point ξ [whose "intensity" is $\delta\,(x - \xi)$]. Then the beam will be deformed in a certain way. We denote the deflection at a point x under the action of unit force applied at the point ξ by $y = G\,(x,\,\xi)$. It is the function $G\,(x,\,\xi)$ that is called the influence function of our problem. We shall show that if the function is known it is easy to determine the deflection under the action of an arbitrary load of intensity $p\,(x)$.

Indeed, let us consider the portion of the load on the infinitesimal segment of the axis from the point ξ to the point $\xi + d\xi$. This load is equal to $p\,(\xi)\,d\xi$. Therefore the deflection under this load at a point x is equal to $G\,(x,\xi)\,p\,(\xi)\,d\xi$ because the linearity we have mentioned above implies that if a load is multiplied by a constant the corresponding deflection is multiplied by the same constant. Adding together all these infinitesimal deflections we obtain the resultant deflection (see Fig. 280):

$$h\,(x) = \int_0^l G\,(x,\,\xi)\,p\,(\xi)\,d\xi \qquad (105)$$

Now we proceed to describe the general scheme for constructing influence functions. Let an external action be applied to an object and let it be described by a function $f(x)$ defined over an interval $a \leqslant x \leqslant b$ (the role of $f(x)$ was played by the function $p(x)$ in our previous example). Let a function $\tilde{f}(x)$, $a \leqslant x \leqslant b$, describe the result of the action (such a result was described by the function $h(x)$ in the example). Thus, every given function f is transformed into a new function $\tilde{f}$. By Sec. XI.6, such a law of transformation of a preimage which is the function f into the image which is the function $\tilde{f}$ is called an **operator**. For instance, the operator of differentiation D transforms functions according to the law $Df = f'$ and thus we have $D(\sin x) = \cos x$, $D(x^3) = 3x^2$ etc. Here $\sin x$ is a preimage which is transformed by the operator D into the image $\cos x$ etc. The concept of an operator is analogous to the concept of a function (see Sec. I.11) but a function transforms numbers into numbers (i.e. the values of an argument into the corresponding values of the function) whereas an operator transforms functions into functions (or, generally, objects of any kind into objects of the same kind or of another kind).

Let us denote the operator which transforms a function $f(x)$ describing an external action into the "response" function $\tilde{f}(x)$ by A, that is $Af = \tilde{f}$. We shall suppose that there is a linearity law here or, as we say, the **superposition principle**: when external actions are added together their results are also added up. This law which can be written in the form

$$A(f_1 + f_2) = Af_1 + Af_2$$

is often applied when the external actions are not very large. An operator possessing this property is said to be a **linear operator** (compare with Sec. XI.6). (Let the reader verify that the operator D is linear.) From the law of linearity we can deduce that if an external action is multiplied by a constant the corresponding result is also multiplied by the constant, that is

$$A(Cf) = CAf$$

where $C = \text{const.}$ (Try to justify this rule first by taking positive integral C and then by putting $C = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots, C = \frac{m}{n}$ where m and n are integers, $C = 0$ and, finally, by passing to negative C.)

We now denote as $G(x, \xi)$ the result of an external action described by the delta function $\delta(x - \xi)$ (regarded as a function of x for every fixed value of ξ), that is

$$A[\delta(x - \xi)] = G(x, \xi)$$

Any function $f(x)$ can be represented as a sum of "impulse functions" of the form depicted in Fig. 281 below (the corresponding summation is depicted in Fig. 281 above). Each of these functions has a singularity only at one point (when we pass to the limit as $d\xi \to 0$) and is therefore equal to $f(\xi)\,d\xi\,\delta(x - \xi)$ (why is it so?). Thus we have

$$f(x) = \sum f(\xi)\,d\xi\,\delta(x - \xi)$$

This is in fact formula (102) written in another form. It follows that

$$A[f(x)] = A\left[\sum f(\xi)\,d\xi\,\delta(x - \xi)\right] = \sum A[f(\xi)\,d\xi\,\delta(x - \xi)] =$$

$$= \sum f(\xi)\,d\xi\,A[\delta(x - \xi)] = \sum f(\xi)\,d\xi\,G(x, \xi)$$

But when $d\xi$ is infinitesimal this sum turns into an integral, and finally we obtain

$$A[f(x)] = \int_a^b G(x, \xi)\,f(\xi)\,d\xi \tag{106}$$

[compare with formula (105)].

An influence function can be determined theoretically in simpler cases (for instance, see the end of Sec. XV.16). In more complicated

cases it can be found experimentally by performing necessary measurements (for example, we can measure the deformation of an elastic system caused by the action of unit force). The essential thing is to verify the linearity of the system in question, that is to find out whether the superposition principle is applicable. The applicability can be deduced theoretically or confirmed by an experiment. Of course, not all systems are linear or even approximately linear. There are such systems that are essentially non-linear, and linear methods are inapplicable to such systems.

It should be noted that the functions f and $\widetilde{f}$ can be defined over different intervals. Moreover, the independent variables x and ξ entering into formula (106)
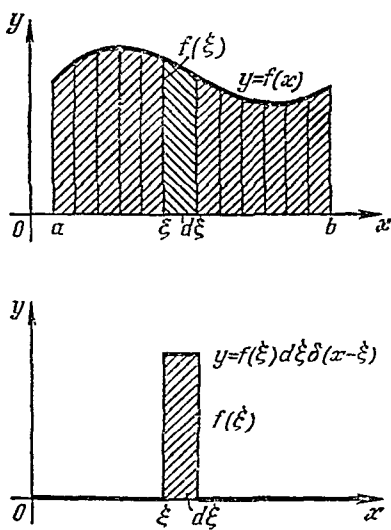
Fig. 281

may have different physical meanings. The independent variable ξ is sometimes interpreted as time. In this case the influence function describes the result of applying unit impulse at the moment ξ.

t sometimes happens that a system in question is linear only
infinitesimal external actions. But an action whose density is
cribed by a delta function cannot be considered to be small.
en the influence function can be defined by the formula

$$G(x, \xi) = \lim_{P \to 0} \frac{1}{P} A[P\delta(x - \xi)]$$

r instance, in our previous example we can find the deflection
responding to a small force P and then divide the result by P.
such circumstances formula (106) is applicable only to the case
small (or, more precisely, infinitesimal) external actions $f(x)$.
27. **Other Generalized Functions.** We now take the approximate
resentation $\delta_N(x)$ of the delta function given in Sec. 25 for
continuous model and differentiate it. We obtain an approximate
resentation of the derivative $\delta'(x)$. It can be seen that $\delta'(x)$
es on the values of both signs and has singularities that are
ore acute" than those of the delta function $\delta(x)$.
We have shown that the δ-function describes the density of
it charge located at the origin (see Sec. 25). Its derivative
(x) describes the density of a *dipole* placed at the same point.
fact, we can obtain such a dipole if we put the charges $-q$ and q
the points $x = 0$ and $x = l$, respectively, and then pass to the
ait, as l tends to zero, retaining the constant value of the quantity
$= ql$ (the *dipole moment*). Thus we obtain two infinitely large
arges of opposite signs with an infinitesimal distance between
em. Before the passage to the limit the density of the charges has
a form

$$q\delta(x - l) - q\delta(x) = -p \frac{\delta(\dot{x} - l) - \delta(x)}{-l}$$

d therefore, after the passage to the limit for $l \to 0$, the density
comes equal to $-p\delta'(x)$.
Integrals involving $\delta'(x)$ are evaluated with the help of integra-
on by parts. For instance, if $f(x)$ has a continuous derivative and
$\alpha < a < \beta$ then

$$f(x)\delta'(x - a)\, dx = f(x)\delta(x - a)\Big|_{x=\alpha}^{x=\beta} - \int_{\alpha}^{\beta} \delta(x - a) f'(x)\, dx = -f'(a)$$

ere we also give an example of incorrect calculations:

$$\int_{\alpha}^{\beta} f(x)\delta'(x - a)\, dx = \int_{a-0}^{a+0} f(x)\delta'(x - a)\, dx =$$

$$= f(a) \int_{a-0}^{a+0} \delta'(x - a)\, dx = f(a)\delta(x - a)\Big|_{x=a-0}^{x=a+0} = 0$$

his is wrong because $\delta'(x - a)$ has a very "acute" singularity and
erefore the substitution of $f(a)$ for $f(x)$ results in an approximation

which is not sufficiently accurate. The correct evaluation of the integral must be performed as above.

Generalized functions can be classified according to the number of successive integrations which must be performed in order to obtain a continuous function. If we assume this classification then continuous functions can be regarded as generalized functions of zero order; functions with finite jump discontinuities and ordinary functions with integrable singularities placed at· finite distances (for example, from the origin) described in Sec. 16 are generalized functions of the first order. The function $\delta\ (x)$ is the simplest example of a generalized function of the second order (see Sec. 25) whereas $\delta'\ (x)$ is of the third order and so on. The differentiation of a function increases the order by unity and the integration reduces it by unity.

When a function with non-integrable singularities is interpreted as a generalized function we must indicate the function of order 0 or 1 from which the function in question can be obtained by differentiation because there can be many such functions. For instance, the function $\frac{1}{x}$ having a non-integrable singularity at $x = 0$ can be regarded, for $x \neq 0$, as being equal to

$$(\ln\ |\ x\ |)'\quad \text{or to}\quad (\ln\ |\ x\ | + e\ (x))' \qquad (107)$$

where $e\ (x)$ is the step function (see Sec. 26) because we have $e'\ (x) = = 0$ for $x \neq 0$. But functions (107) differ by $e'\ (x) = \delta\ (x)$ and their properties are different. In the theory of generalized functions it is preferable to use form (107) for designating the function $\frac{1}{x}$ because after one of these formulas has been chosen (or some other formula of this kind) the function $\frac{1}{x}$ is represented in a unique manner as a generalized function [of course, formulas (107) yield different representations]. Similarly, the function $\frac{1}{x^2}$ can be regarded as $-(\ln\ |\ x\ |)''$ and so on. Functions whose rate of growth at their points of discontinuity is greater than the rate of growth of any power function (for instance, the function $e^{\frac{1}{|x|}}$) should be excluded from this classification. By the way, these functions are not of great importance for applications. It should also be noted that the growth of a function for $x \rightarrow \pm\infty$ is not restricted here.

It turns out that the use of generalized functions makes it possible to extend most of the rules connected with differentiation of different formulas without usual restrictions imposed on the behaviour of the functions. For instance, the Leibniz formula for differentiation of an integral with respect to a parameter (see Sec. 20) becomes true for any type of convergence of improper integrals in question and the like.

'

Differential Equations

A differential equation is an equation connecting two or more functionally dependent variables and their differentials or, which is the same, their derivatives. The problem of forming and solving these equations is widely encountered in physics and engineering. The process of solving a differential equation is called *integration of the differential equation*.

§ 1. General Notions

1. Examples. We have already dealt with some simple differential equations in our course. For instance, take equations (XIV.22). If the force F (s) varies according to Newton's law $\left(\text{i.e. } F = \frac{k}{s^2} \right.$; see Sec. XIV.14$\left.\right)$ we can rewrite the equation in the form

$$dA = \frac{k}{s^2} ds \quad \text{or} \quad \frac{dA}{ds} = \frac{k}{s^2} \tag{1}$$

where the work $A = A$ (s) is an unknown function of the displacement s. Equation (1) is a differential equation, and the unknown function A (s) is found by means of integration.

Another example is equation (XIV.27) which can be rewritten as

$$\frac{dh}{dt} = -\frac{\sigma}{S} \sqrt{2gh} \tag{2}$$

where $h = h$ (t) is an unknown function.

In addition, let us consider an example of elastic vibrations of a material point of mass M about an equilibrium position (see Fig. 282). Here the unknown function $y = y$ (t) expresses the law of vibrations. For simplicity's sake, let us suppose that there is a linear law of elasticity here, that is the elastic force is directly proportional to the deviation of the point from the equilibrium position. Then the force is equal to

$$F = -ky$$

32—0141

where k is the stiffness factor. If there are no other forces then, according to Newton's second law, we have

$$M \frac{d^2y}{dt^2} = F = -ky \tag{3}$$

Thus, the differential equation of the law of vibrations is of the form

$$M \frac{d^2y}{dt^2} + ky = 0 \tag{4}$$

where $y = y(t)$ is the unknown function.

A differential equation of a problem in physics or engineering is always deduced on the basis of a certain law describing a relationship between infinitesimal variations of quantities in question (a differential law). After the differential equation has been integrated we get an integral law describing finite variations of the quantities. The deduction of basic differential equations in a certain branch of science is a very important operation because it essentially determines the course of further development of this branch.
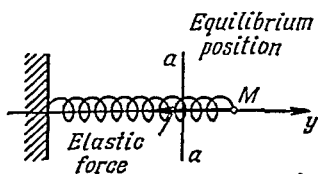
Fig. 282

2. Basic Definitions. A differential equation is usually taken in a form which connects an argument (or several arguments) and an unknown function (or several unknown functions) with its derivatives. Even if we originally have a relationship between differentials it is possible to transform it into a relationship between the derivatives [see formulas (1)]. If the unknown function in a differential equation depends on one variable the differential equation is called **ordinary** (for example, the equations in Sec. 1). If otherwise the equation is called a **partial differential equation** (think why it is called so). In this chapter we shall deal only with ordinary differential equations.

The highest order of the derivative of the unknown function entering into an equation is called the **order of the differential equation.** Thus, equations (1) and (2) are first-order equations whereas equation (4) is a second-order equation. The general form of a differential equation of the nth order is

$$F(x, y, y', y'', \ldots, y^{(n)}) = 0 \tag{5}$$

where $y = y(x)$ is the sought-for function. In particular cases a function F may not depend on some of the quantities entering into (5). For example, equation (4) does not contain the independent variable and the derivative of the first order.

A function is called a **solution** of a differential equation if it reduces the equation to an identity when substituted into the equation.

Even the simplest examples indicate that a differential equation has infinitely many solutions. For instance, taking a simple equation of the form

$$y' = x^2, \qquad y = y(x) \tag{6}$$

we immediately find, by integrating, that

$$y = \frac{x^3}{3} + C \tag{7}$$

This is the **general solution** of equation (6). It contains an **arbitrary constant** C and is the set of solutions containing all solutions of the equation. Making the arbitrary constant assume concrete numerical values we obtain **particular solutions** of equation (6):

$$y = \frac{x^3}{3}, \quad y = \frac{x^3}{3} + 6, \quad y = \frac{x^3}{3} - \frac{\sqrt{2}}{3} \quad \text{etc.}$$

If we take an nth-order equation of the form

$$y^{(n)} = x^2, \quad y = y(x)$$

then its general solution can be found by means of n subsequent integrations and therefore it contains n arbitrary constants. Similarly, the general solution of an equation of the form (5) also contains n arbitrary constants, i.e. it has the form

$$y = y(x, C_1, C_2, \ldots, C_n) \tag{8}$$

We often obtain the general solution in an implicit form

$$\Phi(x, y, C_1, C_2, \ldots, C_n) = 0 \tag{9}$$

Relations (8) and (9) are also called **general integrals** of equation (5). Particular solution can be obtained from (8) or (9) if we make each of the arbitrary constants $C_1, C_2, \ldots, C_n$ take on a certain concrete numerical value. The graph of every particular solution is called an **integral curve** of the differential equation in question. Substituting these concrete numbers for the constants $C_1, C_2, \ldots \ldots, C_n$ into equation (8) or (9) we get an equation of the integral curve.

To isolate a unique particular solution from the general solution we must set some additional conditions. Such conditions are often taken as so-called **initial conditions**. If we have a process developing in time then such conditions are a mathematical expression of the initial state of the process.

For example, take the process of vibrations considered in Sec. 1. The physical meaning of the problem makes it clear that a particular (concrete) vibration will be completely specified if we set the values

ιe initial deviation of the point from the equilibrium position
of the initial velocity. Therefore the initial conditions for
tion (3) are of the form

$$y = y_0 \quad \text{and} \quad \frac{dy}{dt} = v_0 \quad \text{for} \quad t = t_0 \tag{10}$$

e y_0 and v_0 are the given values. In the general case of an equa-
of form (5) the initial conditions are

$$y_0, \quad y' = (y')_0, \ldots, \quad y^{(n-1)} = (y^{(n-1)})_0 \quad \text{for} \quad x = x_0 \tag{11}$$

e the values y_0, $(y')_0$, ..., $(y^{(n-1)})_0$ are given. The general
ion [for instance, of form (9)] containing n arbitrary constants.
possible (at least theoretically) to determine the values of the
tants taking advantage of the n relations we have set. Thus,
rally speaking, the number of additional conditions (11) is
ient for determining the arbitrary constants and specifying
)articular solution. It appears natural, from the physical point
iew, that if a differential law controlling the development of
)cess and an initial state of the process are given then the pro-
itself is completely specified.
ndition (11) for an equation of the first order of form (6) means
for a certain value $x = x_0$ we must assign a value $y = y_0$.
instance, let it be necessary to isolate a solution for which
$= 2$. Then (7) implies $2 = \frac{1^3}{3} + C$, i.e. $C = \frac{5}{3}$. Hence, the
ιt-for particular solution has the form

$$y = \frac{x^3 + 5}{3}$$

e problem of finding a particular solution of a differential
tion when certain initial conditions are given is called the
hy problem (initial-value problem).
we shall see in Sec. 7, there are some differential equations
h possess so-called **singular solutions** in addition to the parti-
ʹ ones contained in the general solution, that is solutions that
)t enter into the general solution.

§ 2. First-Order Differential Equations

Geometric Meaning. The general form of a first-order diffe-
al equation can be written as

$$F(x, y, y') = 0 \tag{12}$$

e $y = y(x)$ is the unknown function. For simplicity's sake,
rst suppose that the equation is solved for the derivative of the

unknown function. Then the equation takes the form

$$y' = f(x, y) \tag{13}$$

According to Sec. 2, the initial condition for equation (12) must be written as

$$y = y_0 \quad \text{is given for} \quad x = x_0 \tag{14}$$

To get a geometric interpretation of equation (13) let us introduce a plane with Cartesian coordinates x and y. Then every particular solution is represented by a curve (the integral curve) lying in the x, y-plane, these curves being yet unknown. But if we take an arbitrary point $M(x, y)$ in the plane we can compute the value
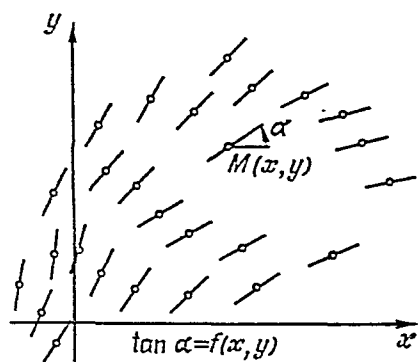
Fig. 283
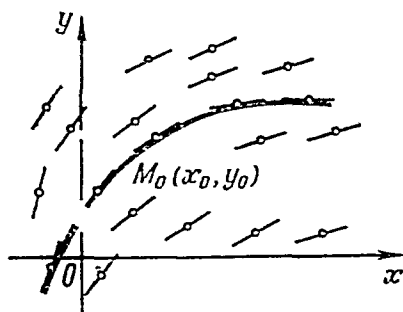
Fig. 284

of $f(x, y)$ which, as it is prescribed by equation (13), must equal the slope of the tangent (see Sec. IV.3) to the desired curve at the point $M(x, y)$ provided the curve passes through the point M.

We can therefore perform the following procedure: let us draw (mentally) a small line segment with the slope $\tan \alpha = f(x, y)$ at each point $M(x, y)$ (see Fig. 283). Practically we can draw only a number of such segments but theoretically we can regard the segments as drawn through all the points. Thus we obtain the so-called direction field in the plane defined by equation (13) (the general concept of a field was introduced in Sec. IX.9). Hence, we see that the integral curves of equation (13) must pass through the points $M(x, y)$ of the x, y-plane in such a way that each curve should touch the segment at each point it passes through.

Thus, equation (13) defines a direction field in the x, y-plane. On the other hand, initial condition (14) defines a point $M_0(x_0, y_0)$ through which the desired integral curve should pass. From the geometric point of view it is clear that the above condition completely specifies the integral curve (see Fig. 284). In other words, initial condition (14) being given, equation (13) has a completely

specified unique solution. A more detailed investigation carried out by Cauchy shows that the above assertion holds provided the function f is continuous at the point M_0 and has a finite value of its derivative $\frac{\partial f}{\partial y}$ at the point (Cauchy's conditions are in fact sufficient for the existence and uniqueness of the solution but they are not necessary; certain cases when the above conditions of *Cauchy's theorem* are not fulfilled will be considered in Sec. 7; as it will be seen, this may yield non-fulfillment of the uniqueness of a solution).

Fig. 285

PR is the direction of the field

These considerations concerning a direction field can be illustrated by means of the well-known experiment with iron filings placed in a magnetic field. The arrangement of the filings demonstrates a direction field whose integral curves are the so-called magnetic lines of force.

The geometric meaning of equation (13) we have just discussed enables us to construct (approximately) the integral curves of the equation. To do this we depict the directions of the field at as many points as possible and then draw the curves according to the directions.

Practically, in constructing a direction field it is more convenient to take the points belonging to the so-called isoclines instead of choosing the points arbitrarily. An isocline is a locus of points at which the field has the same direction. We can derive an equation of an isocline if we equate the right-hand side of equation (13) to a constant. Thus we write

$$f(x, y) = k$$

where k is the slope of the field corresponding to the isocline we have taken.

For instance, let us take the equation $y' = x + y$.

Equating the right-hand side to the constants $-2, -\frac{3}{2}, -1, -\frac{1}{2}$, 0, 1 and 2 we obtain the corresponding isoclines which, in this case, are straight lines (i.e. the lines $x + y = -2$ etc.). These isoclines are depicted in Fig. 285. The direction of the field is indicated on

each of the isoclines. To find the direction of the field on a given isocline corresponding to the slope k of the field we can construct the rectangular triangle PQR having the base $PQ = 1$ parallel to the x-axis and the altitude $QR = k$. Then the side PR will indicate the desired direction. There are also several integral curves in Fig. 285 which are drawn in accordance with the directions. We see that the straight line $x + y = -1$ is one of these curves. We also see that the locus of the lowest points of the integral curves is the straight line $x + y = 0$. In the case of an equation of general form (13) in order to find the loci of the highest and of the lowest points belonging to the integral curves it is necessary to construct the isocline $f(x, y) = 0$. (Think how we can find the locus of the points of inflection of integral curves in the general case.)

4. **Integrable Types of Equations.** We say that a differential equation is integrable by quadratures if its general solution is expressible in an explicit or implicit form which may contain quadratures (i.e. indefinite integrals) of some known functions. We consider the integration completed even if these quadratures are not in fact computed (the theory of the integral given in Chapters XIII and XIV deals with the methods of integrating functions). As we know, a quadrature may not be expressible in terms of elementary functions but nevertheless in this case we shall as well consider the integration of a differential equation completed. Unfortunately, even most of the simplest equations are not integrable by quadratures and it is therefore necessary to investigate them by means of some other methods which will be discussed later. But there exist certain classes of differential equations that are integrable by quadratures and we shall study them here.

1. *Differential equations with variables separable* have been already dealt with (see Sec. XIV.7). They have the general form

$$\frac{dy}{dx} = f(x)\,\varphi(y) \tag{15}$$

and their general solution is expressed by the formula

$$\int \frac{dy}{\varphi(y)} = \int f(x)\,dx + C \tag{16}$$

We have written the arbitrary constant C (which we usually regard as being included into the sign of indefinite integral) in order to stress that the constant enters into the general solution. Thus, we see that equation (15) has been integrated by quadratures [this is expressed by formula (16)].

As an example, take a simple equation of the form

$$\frac{dy}{dx} = 2xy$$

Separating the variables and performing integration we obtain

$$\frac{dy}{y} = 2x\,dx, \quad \ln|y| = x^2 + C, \quad |y| = e^{x^2 + C}, \quad y = \pm e^C e^{x^2} \quad (17)$$

The answer can be written in a different form if we notice that the expression $\pm e^C$ can also be regarded as an arbitrary constant. Therefore $y = Ce^{x^2}$. It is apparent that the symbol C in the last formula designates a new constant which is different from the one entering into formula (17).

To avoid the change of notation we could write

$$\ln|y| = x^2 + \ln C$$

while integrating equation (17), since $\ln C$ is also an arbitrary constant. Then, raising, we get

$$|y| = Ce^{x^2}, \quad y = \pm Ce^{x^2} \quad \text{or, simply,} \quad y = Ce^{x^2}$$

because the signs $+$ and $-$ may also be included into C, that is C in the expression $y = Ce^{x^2}$ is allowed to take on values of arbitrary signs. Further we are going to perform transformations of this kind without specific stipulations.

We can similarly construct the general solution of the equation

$$P(x)\,Q(y)\,dx + R(x)\,S(y)\,dy = 0$$

which is also an equation with variables separable.

2. *Equations homogeneous in the argument and in the unknown function.* An equation of form (12) is said to be homogeneous in x and y if its left-hand side is a homogeneous function in x and y, that is if

$$F(tx,\,ty,\,y') \equiv t^k F(x,\,y,\,y')$$

(the general definition of a homogeneous function was given in Sec. IX.12).

Then equation (12) can be rewritten in the form

$$F\left(x\cdot 1,\,x\cdot\frac{y}{x},\,y'\right) = 0, \quad x^k F\left(1,\,\frac{y}{x},\,y'\right) = 0, \quad \text{i.e.} \ F\left(1,\frac{y}{x},\,y'\right) = 0$$

Solving the last equation for y' we obtain

$$y' = \varphi\left(\frac{y}{x}\right) \qquad\qquad (18)$$

This equation is easily integrated by means of the substitution

$$\frac{y}{x} = u, \quad y = ux, \quad y' = u'x + u$$

where $u = u(x)$ is a new unknown function which replaces y. Substituting $y = ux$ into (18) we derive

$$u'x + u = \varphi(u), \quad \frac{du}{dx}\,x = \varphi(u) - u, \quad \frac{du}{\varphi(u) - u} = \frac{dx}{x}$$

and thus the variables have been separated. To complete the integration we must solve the last equation and then return from u to the original unknown function y.

Let us take a more general equation than (18), namely,

$$\frac{dy}{dx} = \varphi \left(\frac{ax+by+c}{mx+ny+p} \right)$$

Suppose that the binomials $ax + by$ and $mx + ny$ are not proportional to each other. Then we can make a substitution of the form $x = x_1 + \alpha$ and $y = y_1 + \beta$ where the parameters α and β should be chosen so that there should be no absolute terms in the denominator of the resulting fraction. After that we substitute u for the ratio $\frac{y_1}{x_1}$ and thus obtain an equation with variables separable.

3. *Linear equations.* An equation of form (12) is called linear if its left-hand side is a linear function in the unknown function and its derivative, i.e. an equation of the form

$$a(x) y' + b(x) y + c(x) = 0$$

Dividing the equation by $a(x)$ we obtain

$$y' + p(x) y = f(x) \tag{19}$$

where $p = \dfrac{b}{a}$ and $f = -\dfrac{c}{a}$.

Equation (19) is called a **homogeneous** linear equation if $f(x)$ is identically equal to zero. If otherwise the equation is called non-homogeneous. To solve the general non-homogeneous equation of form (19) we first investigate the auxiliary homogeneous equation

$$z' + p(x) z = 0 \tag{20}$$

which corresponds to (19) and is obtained from (19) by dropping the non-homogeneity term $f(x)$. The variables in equation (20) are easily separated:

$$\frac{dz}{dx} = -p(x) z, \quad \frac{dz}{z} = -p(x) dx, \quad \ln|z| = -\int p(x) dx + \ln C,$$

$$z = C \exp \left[-\int p(x) dx \right] = C z_1 \tag{21}$$

where z_1 is obtained from z if we substitute $C = 1$. Thus, z_1 is a particular solution of equation (20).

After z has been found we seek a solution of equation (19) in the form

$$y = \varphi(x) z_1 \tag{22}$$

where z_1 is the same as in formula (21) and $\varphi(x)$ is a function yet unknown. Such a replacement of the former arbitrary constant entering into formula (21) by a function entering into formula (22)

is an application of a general method called the **method of variation of arbitrary constants (parameters)** which was introduced by Lagrange.

Substituting (22) into equation (19) we receive

$$\varphi' z_1 + \varphi z_1' + p\varphi z_1 = f, \quad \varphi' z_1 + \varphi (z_1' + p z_1) = f$$

The function z_1 satisfying equation (20), the expression inside the brackets equals zero. Therefore,

$$\varphi'(x) = \frac{f(x)}{z_1(x)}, \quad \varphi(x) = \int \frac{f(x)}{z_1(x)} \, dx + C,$$

$$y = z_1(x) \int \frac{f(x)}{z_1(x)} \, dx + C z_1(x)$$

The last expression is the general solution of equation (19). The first term of the expression can be obtained by substituting $C = 0$. Hence, the first term is a particular solution of the equation.

Thus, *the general solution of non-homogeneous linear equation* (19) *is equal to the sum of a particular solution of the non-homogeneous equation and the general solution of the corresponding homogeneous equation*. The equation

$$y' + p(x) y = f(x) y^n$$

is called **Bernoulli's equation**. It can be reduced to a linear equation by means of dividing both sides by y^n and introducing the change of variables of the form $y^{1-n} = u$ (check it up!).

There are some other types of equation integrable by quadratures (see [24]) but most of the differential equations are not integrable by quadratures. For example, in the general case we cannot integrate by quadratures the equation $y' = y^2 + f(x)$ which is called **Riccati's equation** (named after J. Riccati, 1676-1754, an Italian mathematician). Riccati's equations are applied to some problems. There are many other simple differential equations that are not integrable by quadratures.

5. Equation for Exponential Function. Let us consider the equation

$$\frac{dy}{dx} = ky \quad (k = \text{const}) \tag{23}$$

which indicates that the rate of change of the quantity y related to the quantity x is proportional to the current value of y. Hence, if $y(x) > 0$ then y increases for $k > 0$ and decreases for $k < 0$. Such a simple relationship between a quantity and its rate of change is often used as a first approximation in investigating various processes.

The variables in equation (23) can be separated which yields

$$\frac{dy}{y} = k \, dx, \quad \ln|y| = kx + \ln C, \quad y = C e^{kx}$$

In addition, if there is an initial condition $y(x_0) = y_0$ we obtain

$$y_0 = Ce^{kx_0}, \quad C = y_0 e^{-kx_0}, \quad \text{i.e.} \quad y = y_0 e^{k(x-x_0)} \qquad (24)$$

Thus, the general solution of equation (23) is an exponential function (see Sec. I.27). It is characteristic of this solution that if we make x assume the values forming an arithmetic progression with common difference Δx then the corresponding values of y form a geometric progression with common ratio $e^{k\Delta x}$. We can easily find the value of Δx for which y increases or decreases twice for every step Δx. Indeed, this being so, we must have

$$|k\Delta x| = \ln 2, \quad \text{i.e.} \quad \Delta x = \frac{\ln 2}{|k|} \qquad (25)$$

If $k > 0$ and $y_0 > 0$ formula (24) describes the so-called law of exponential growth of the quantity y which is characteristic of different *chain reactions*. As an example, let us consider the process of reproduction of bacteria in a culture medium. when their number is not too large. We suppose that all the bacteria reproduce more or less independently. Then we see that the rate of growth of the number u of the bacteria measured in certain units is proportional to the number, i.e.

$$\frac{du}{dt} = ku, \quad u = u_0 e^{k(t-t_0)}$$

There are many problems of this kind that can be investigated in a similar way, the problem of calculating the growth of a capital deposited in a bank being one of them.

If $k < 0$ then formula (24) expresses an exponential decrease of a quantity y. For example, this is the case when we investigate the process of a radioactive decay. In fact, let us denote the mass of a remaining (not disintegrated) radioactive substance by m. If we suppose that different parts of the mass are decaying independently then we conclude that the rate of decay of the mass is proportional to the current value of the mass, that is

$$\frac{dm}{dt} = -pm, \quad m = m_0 e^{-p(t-t_0)}$$

In particular, note that after the elapse of time $\Delta t = \dfrac{\ln 2}{p}$ the value of m becomes half as large, this being suggested by formula (25). The time interval Δt is known as the **half-life period** (or simply half-life of the radioactive substance). For instance, Δt is approximately equal to 1.8×10^3 years for radium. This means that if an initial mass of radium is not replenished then in $1.8 \quad 10^3$ years we shall have half of the initial quantity and after another $1.8 \quad 10^3$ year period passes we shall have a quarter of the initial mass etc.

There are many phenomena, such as the decrease of the atmospheric pressure with the growth of the altitude or the discharge of a capacitor through a resistance, which are investigated in a similar way.

The equation of a problem can sometimes be transformed to form (23) by means of some simple techniques. For example, as it was shown in Sec. VIII.7, the electric current flow i in a circuit consisting of a resistance R and an inductance L satisfies the equation

$$L\frac{di}{dt} + Ri = u \qquad (26)$$

when a constant voltage u is applied to the terminals of the chain.

Equation (26) is a non-homogeneous linear equation which can be integrated (solved) by means of the method described in Sec. 4. But it is easier to transform the equation in the following way:

$$L\frac{di}{dt} = -Ri + u = -R\left(i - \frac{u}{R}\right), \quad \frac{d\left(i - \frac{u}{R}\right)}{dt} = -\frac{R}{L}\left(i - \frac{u}{R}\right)$$

whence

$$i - \frac{u}{R} = \left(i_0 - \frac{u}{R}\right)e^{-\frac{R}{L}(t-t_0)}, \quad i = \frac{u}{R} + \left(i_0 - \frac{u}{R}\right)e^{-\frac{R}{L}(t-t_0)}$$

We obtain a simpler case when there is no initial current in the circuit. Indeed, let $t_0 = 0$ be the initial moment. Then $u(t_0) = u(0) = u_0 = 0$ and we receive the formula

$$i = \frac{u}{R}\left(1 - e^{-\frac{R}{L}t}\right) \qquad (27)$$

The graph of relationship (27) is shown in Fig. 286. We see that the current increases and exponentially tends to the limiting steady-state value $\frac{u}{R}$, as $t \to \infty$. This value can also be easily found from equation (26) if we take into account that $\frac{di}{dt} \to 0$ as $t \to \infty$ in the process of current rise. Therefore, in the limit we have $Ri = u$, i.e. $i = \frac{u}{R}$. Thus, when the current becomes practically steady-state the whole voltage drop is on the resistance. During the time period

$$\tau = \frac{\ln 2}{\dfrac{R}{L}} = \frac{L}{R}\ln 2$$

the deviation of the current flow from its limit value becomes twice as small.

The fact that it is the constant e that appears as the base of the exponential function in formula (24) is one of the main reasons which account for the important role of the constant in mathematics and its applications.

6. Integrating Exact Differential Equations. A differential equation of the first order is often written in the *symmetric form*

$$P\ (x,\ y)\ dx + Q\ (x,\ y)\ dy = 0 \qquad (28)$$

in place of form (13). Here $P\ (x,\ y)$ and $Q\ (x,\ y)$ are given functions, and it is the functional relationship between x and y that is considered to be unknown. We can easily pass from one form to another.
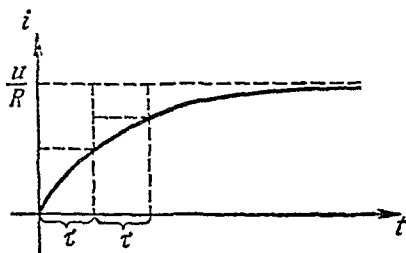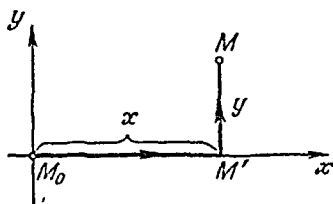
Fig. 286   Fig. 287

For instance, to transform equation (28) to form (13) we must divide both sides of (28) by $Q\ dx$ and then transpose $\dfrac{P}{Q}$ to the right-hand side. Form (28) is preferable in those cases when the variables x and y are regarded as being equivalent, that is when we do not set beforehand which of the variables x and y is an argument and which is a function.

There is a special case here when the left-hand side of equation (28) is the total (exact) differential of a function, that is

$$P\ dx + Q\ dy \equiv du\ (x,\ y) \qquad (29)$$

Then we can easily integrate the equation. Actually, in this case the equation can be rewritten as $du = 0$ and hence, integrating, we get the general solution

$$u\ (x,\ y) = C \qquad (30)$$

where C is an arbitrary constant, as usual.

At the end of Sec. XIV.24 we obtained a condition guaranteeing the existence of such a function $u\ (x,\ y)$, namely, the condition

$$\frac{\partial P}{\partial y} \equiv \frac{\partial Q}{\partial x} \qquad (31)$$

This condition is necessary and sufficient for the left-hand side of equation (28) to be an exact differential. The function $u\ (x,\ y)$

is found according to formula (XIV.97) in which, of course, we must drop the last summand under the integral sign in our case. As was pointed out in Sec. XIV.24, generally we can arrive at a multiple-valued function u if the domain under consideration is multiply-connected. But even in this case formula (30) expresses the general solution of equation (28).

As an example, let us take the equation

$$(x^2 + 2xy)\,dx + (x^2 - y^3)\,dy = 0 \qquad (32)$$

Here we have

$$\frac{\partial P}{\partial y} = \frac{\partial\,(x^2 + 2xy)}{\partial y} = 2x \quad \text{and} \quad \frac{\partial Q}{\partial x} = \frac{\partial\,(x^2 - y^3)}{\partial x} = 2x$$

Thus, condition (31) is fulfilled. In order to apply formula (XIV.97) to constructing the function u we choose the point M_0 at the origin of coordinates, for definiteness. We also choose the path connecting M_0 with the variable point $M\,(x,\,y)$ in the way shown in Fig. 287. Consequently, we obtain

$$u\,(x,\,y) = \int\limits_{M_0M} [(x^2 + 2xy)\,dx + (x^2 - y^3)\,dy] =$$

$$= \int\limits_{M_0M'} [(x^2 + 2xy)\,dx + (x^2 - y^3)\,dy] +$$

$$+ \int\limits_{M'M} [(x^2 + 2xy)\,dx + (x^2 - y^3)\,dy] \qquad (33)$$

We must put $y = 0$ and $dy = 0$ in the first integral and regard x and dx as $x = \text{const}$ and $dx = 0$ in the second integral (why?). From this we receive

$$u\,(x,\,y) = \int\limits_0^x x^2\,dx + \int\limits_0^y (x^2 - y^3)\,dy = \frac{x^3}{3} + x^2y - \frac{y^4}{4} \qquad (34)$$

Hence, the general solution of equation (32) has the form

$$\frac{x^3}{3} + x^2y - \frac{y^4}{4} = C$$

Let us similarly treat the equation

$$-\frac{y\,dx}{x^2 + y^2} + \frac{x\,dy}{x^2 + y^2} = 0 \qquad (35)$$

By the way, the equation can be easily integrated in a direct way because the variables in the equation are separable. The equation makes sense for all values of x and y except $x = 0$, $y = 0$ since both coefficients P and Q are discontinuous at the point $(0, 0)$.

We must therefore take the x, y-plane with the origin of coordinates punctured. A plane with one point punctured is not simply-connected. As it was shown in Sec. XIV.24, it is doubly-connected.

Condition (31) is also fulfilled for equation (35) (check it up!). In constructing the function u in accordance with formula (XIV.97) we can choose the point M_0 anywhere except, of course, the origin of coordinates. For instance, let us take the point M_0 (1, 0). We first consider the case $x > 0$. Performing calculations analogous to (33) and (34) we find $u = \operatorname{arc\,tan} \frac{y}{x}$ (check up the result!). The same function $u = \operatorname{arc\,tan} \frac{y}{x}$ satisfies relation (29) for $x < 0$ too.

But if we consider the function $u = \operatorname{arc\,tan} \frac{y}{x}$ to be defined over the whole x, y-plane (with the point (0, 0) removed) it will be discontinuous on the straight line $x = 0$. To get rid of the discontinuity we can put

$$u = \operatorname{Arc\,tan} \frac{y}{x} = \varphi$$

where φ is the polar angle of the point (x, y). This function is not single-valued. Even if we take a certain point $M \neq 0$ and choose a certain value of φ for the point, the argument φ will gain the increment 2π after M traces a closed path round the origin. But nevertheless the general solution of equation (35) is of the form

$$\operatorname{Arc\,tan} \frac{y}{x} = C, \quad \text{i.e.} \quad \frac{y}{x} = \tan C = C_1 \quad \text{and} \quad y = C_1 x$$

where C_1 is an arbitrary constant. From the geometric point of view we have obtained the totality of all the possible straight lines passing through the origin of coordinates.

It may happen that condition (31) does not hold for equation (28). Then the left-hand side of such an equation is not a total differential. It can be shown that in this case there always exists a factor such that the equation becomes exact after being multiplied by the factor. Then the equation can be integrated provided the factor is known. For example, the left-hand side of the equation $-y\,dx + x\,dy = 0$ does not satisfy condition (31). But if we multiply the equation by the factor $\frac{1}{x^2 + y^2}$ the equation is transformed to form (35), which satisfies condition (31). Generally, such a factor is called the **integrating factor** of equation (28). There are no general techniques for finding an integrating factor but it can be found for certain specific classes of differential equations. Besides, the concept of an integrating factor is used in some theoretical investigations.

7. **Singular Points and Singular Solutions.** There are some cases when a first-order equation written in the form

$$y' = f(x, y) \tag{36}$$

[see equation (13)] or in the form

$$P(x, y)\, dx + Q(x, y)\, dy = 0 \tag{37}$$

[see equation (28)] possesses more than one integral curve passing through some point in the x, y-plane or has no such curves. The points of this kind are called **singular points** of the equation in question. They can either be **isolated** or form entire **singular curves**.

We begin our investigation of equation (36) with a simple special example, namely, with the equation

$$y' = y^\alpha \quad (\alpha > 0) \tag{38}$$

Let us regard $y(x)$ as being non-negative $(y \geqslant 0)$. Equation (38) can be easily integrated:

$$\frac{dy}{y^\alpha} = dx, \; \frac{y^{1-\alpha}}{1-\alpha} = x - C \text{ for } \alpha \neq 1 \text{ and } \ln y = x - C \text{ for } \alpha = 1 \tag{39}$$

Here we have written $-C$ instead of C for the convenience of our further considerations. The sign in front of C does not matter because C itself can have any sign. Let us distinguish between the following two cases.

1. $\alpha > 1$. Then solution (39) can be put down as

$$y = \frac{1}{(\alpha-1)^{\frac{1}{\alpha-1}}} \cdot \frac{1}{(C-x)^{\frac{1}{\alpha-1}}} = \frac{\text{const}}{(C-x)^{\frac{1}{\alpha-1}}}$$

which implies that if x varies from $-\infty$ to C then y increases from zero to infinity. The graph is shifted along the x-axis as the constant C is changed. The family of integral curves thus obtained is depicted in Fig. 288. The x-axis itself is an integral curve. It can be obtained as the limit when $C \to \infty$. We see that in this case, for each point of the upper half-plane, there is one and only one integral curve passing through the point. Our example also shows that the solution $y(x)$ of an equation may not be defined over the whole x-axis and may exist only for a certain part of the axis. Indeed, in this example $y(x)$ exists only on the interval $-\infty < x < C$.

In the case $\alpha = 1$ we also get a similar uniqueness (check up this assertion!).

2. $0 < \alpha < 1$. Then we can write solution (39) in the form

$$y = (1-\alpha)^{\frac{1}{1-\alpha}} (x-C)^{\frac{1}{1-\alpha}} = \text{const}\, (x-C)^{\frac{1}{1-\alpha}} \tag{40}$$

from which it follows that if x varies from C to ∞ the variable y increases from zero to infinity. The corresponding family of integral curves is shown in Fig. 289. The x-axis is an integral curve again, which is evident from equation (38), but now it cannot be obtained from formula (40) for any value of C. In this case, for each point of the x-axis, there are two distinct integral curves passing through the point, namely, the x-axis itself and the corresponding curve defined by formula (40). Recall·that the problem of constructing
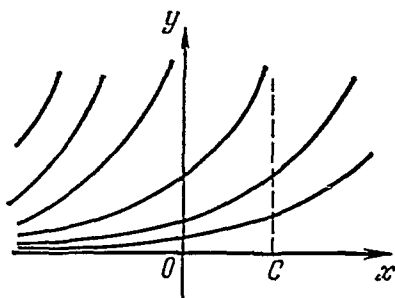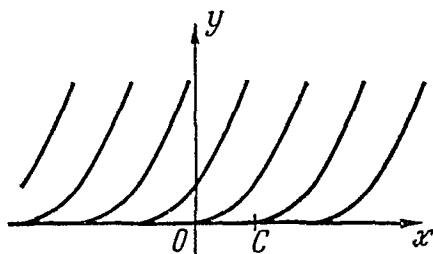
Fig. 288   Fig. 289

an integral curve passing through a given point of the plane is called the Cauchy problem. Thus, in our case the uniqueness of solution of the Cauchy problem is violated.

We see that in the second case the points belonging to the x-axis become singular points. To find out what is the cause of this phenomenon we compute the derivative of the right-hand side of equation (38) at these points, that is for the values $y = 0$. Calculating we obtain

$$\left(\frac{\partial (y^{\alpha})}{\partial y} \right)_{y=0} = (\alpha y^{\alpha-1})_{y=0} = 0 \quad \text{for} \quad \alpha > 1,$$

$$\left(\frac{\partial (y^{\alpha})}{\partial y} \right)_{y=0} = 1 \quad \text{for} \quad \alpha = 1$$

and

$$\left(\frac{\partial (y^{\alpha})}{\partial y} \right)_{y=0} = \infty \quad \text{for} \quad \alpha < 1$$

Therefore, as a point (x, y) approaches the x-axis in the case $0 < \alpha < 1$, the direction of the field changes so fast that every integral curve approaches the axis at a finite point but not at infinity, as we had in Fig. 288.

We see that in the case under consideration the conditions of Cauchy's theorem on existence and uniqueness of the solution are not fulfilled because they include the requirement that the derivative $\frac{\partial f}{\partial y}$ should be finite. In other cases we can also obtain more than

one integral curve passing through a point M_0 if the value of the derivative $\left(\frac{\partial f}{\partial y}\right)_{M_0}$ is infinite at the point M_0 although this is not a necessary consequence of the fact.

In particular, if $\frac{\partial f}{\partial y}$ approaches infinity on a curve (L) and if the curve itself is an integral curve then, as a rule, besides (L), there is at least one more integral curve passing through each point of (L). In this case we say that (L) is a **singular integral curve** which means that (L) is an integral curve whose all points are singular points. The corresponding solution for which a singular integral curve serves as its graph is called a **singular solution**. A singular solution does not usually enter into the general solution, that is, as a rule, it cannot be obtained from the general solution for any value of the arbitrary constant. For instance, the x-axis is a singular integral curve and the function $y \equiv 0$ is the corresponding singular solution of equation (38) in the case $0 < \alpha < 1$ (why?).

There is another approach to the notion of a singular solution. Fig. 289 shows that the x-axis is the envelope of the family of integral curves (see Sec. XII.5) in the case when $0 < \alpha < 1$ for equation (38).

In the general case the envelope of a family of integral curves is also an integral curve because its tangent coincides with the direction of the field at every point provided such an envelope exists. At the same time such an envelope is a singular integral curve since there are other integral curves passing through the points of the envelope. This leads to a method of finding singular solutions based on the consideration given in Sec. XII.5. Suppose we have managed to obtain the general solution in the form $\Phi(x, y, C) = 0$. Then, by Sec. XII.5, we can find a singular solution if we eliminate C from the equations

$$\Phi(x, y, C) = 0 \quad \text{and} \quad \Phi'_C(x, y, C) = 0 \tag{41}$$

[let the reader perform the calculations for solution (40)].

We now proceed to investigate equation (37). For the sake of simplicity, let us suppose that the functions P and Q are continuous and that their derivatives of the first order are finite. Equation (37) can be rewritten in the form

$$\frac{dy}{dx} = -\frac{P(x, y)}{Q(x, y)} \quad \text{or} \quad \frac{dx}{dy} = -\frac{Q(x, y)}{P(x, y)} \tag{42}$$

We can therefore apply above-mentioned Cauchy's theorem concerning equation (36). Thus, if $Q(x_0, y_0) \neq 0$ or $P(x_0, y_0) \neq 0$ at a point $M_0(x_0, y_0)$ then there exists a unique integral curve passing through the point $M_0(x_0, y_0)$. [To apply Cauchy's theorem it is

sufficient to denote one of the right-hand sides (42) which has a non-zero denominator as $f(x, y)$.] But if

$$P(x_0, y_0) = 0, \quad Q(x_0, y_0) = 0 \tag{43}$$

then equation (37) no longer defines a certain relationship between dx and dy at the point $M_0(x_0, y_0)$ and therefore the direction field turns out to be undetermined at the point. Hence, the singular points of equation (37) are defined by relationship (43).
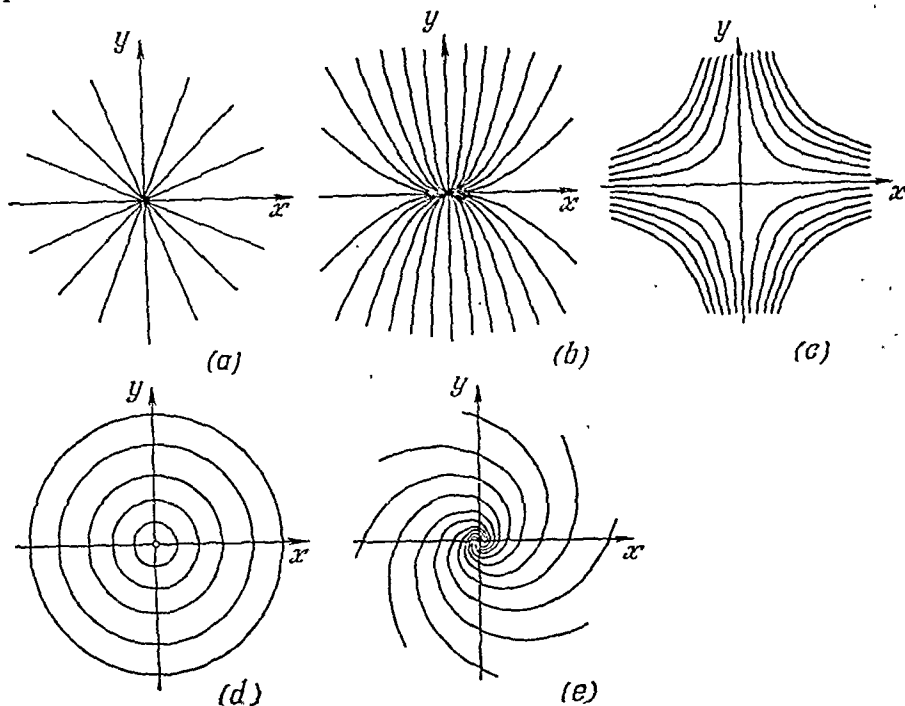
Fig. 290

Singular points of differential equations
(a) Nodal point: $y\, dx - x\, dy = 0$, $y = Cx$
(b) Nodal point: $2y\, dx - x\, dy = 0$, $y = Cx^2$
(c) Saddle point: $y\, dx + x\, dy = 0$, $xy = C$
(d) Centre: $x\, dx + y\, dy = 0$, $x^2 + y^2 = C$
(e) Focal point: $(x + y)\, dx - (x - y)\, dy = 0$, $\rho = Ce^\varphi$
(in polar coordinates)

In Fig. 290 we give some examples of the most widely encountered types of singular point together with their *names*. (Let the reader verify the solutions written in Fig. 290 and the forms of their graphs which are also shown there.) The origin of coordinates is the only singular point in all these examples. In examples (a), (b) and (e) there are infinitely many integral curves passing through the singular point; there are two *such curves in* example (c) whereas there are no integral curves passing through the point in example (d). It

should be noted that the coordinate axes themselves are integral curves in examples (a), (b) and (c). To integrate equation (e) in Fig. 290 it is convenient to transform the equation to polar coordinates.

8. Equations Not Solved for the Derivative. The equation

$$F (x, y, y') = 0 \tag{44}$$

differs from equation (13) investigated in Sec. 3 because in this case y' is an implicit function of x and y. A characteristic feature of implicit functions is that generally they are many-valued (see Sec. I.20). Therefore, if we solve equation (44) for y' (which is theoretically possible but may be difficult to realize practically) we shall get several solutions in the general case:

$$y' = f_1 (x, y), \quad y' = f_2 (x, y), \ldots$$
$$\ldots, y' = f_k (x, y) \tag{45}$$

Fig. 291

Each of the solutions satisfies equation (44).

Each of equations (45) defines its own direction field in the plane and generates a family of integral curves which cover the plane (see Sec. 3). Therefore if in a certain part of the x, y-plane equation (44) possesses k solutions with respect to y' we have the superposition of these k direction fields. Hence, there are k integral curves passing through each point of the part of the plane, that is an initial condition $y (x_0) = y_0$ defines k solutions (see Fig. 291 where we have taken $k = 3$).

In Sec. 7 we considered singular points, singular curves and singular integral curves of equation (36), and equation (44) may likewise have them. Equation (44) suggests that

$$\frac{\partial y'}{\partial y} = - \frac{F'_y (x, y, y')}{F'_{y'} (x, y, y')}$$

(see Sec. IX.13) and therefore, by Cauchy's theorem (see Sec. 3), such points and curves may appear only if equation (44) and the equation

$$F'_{y'} (x, y, y') = 0 \tag{46}$$

hold simultaneously (of course, if $F'_y \neq \infty$).

In particular, a singular solution (provided it exists) whose graph is the envelope of a family of integral curves can be obtained by
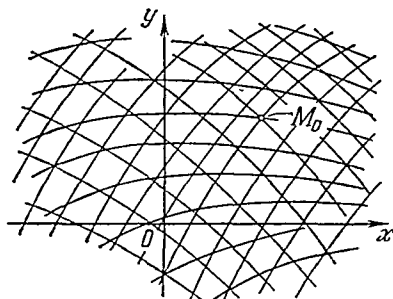
eliminating y' from (44) and (46). There is another method of constructing a singular solution based on formulas (41).

9. Method of Integration by Means of Differentiation. There are certain cases when equation (44) can be integrated if it is differentiated beforehand. For instance, let us take the equation

$$x = f(y') \qquad (47)$$

Such equations are usually written in the form

$$x = f(p) \qquad (48)$$

where $p = y'$.

Differentiating both sides we derive

$$dx = f'(p)\, dp$$

By means of the last equality and formula $\frac{dy}{dx} = p$ we find the following expression of dy:

$$dy = p\, dx = pf'(p)\, dp$$

This implies

$$y = \int pf'(p)\, dp + C \qquad (49)$$

Equalities (48) and (49) simultaneously define a functional relation between x and y which is expressed parametrically (see Sec. II.6), the variable p being the parameter. Thus we have obtained the general solution of equation (47) in a parametric form. An equation of the form $y = f(y')$ can be solved in a similar way (verify it!).

The so-called **Lagrange's equation** of the form

$$y = f(y')\, x + g(y'), \quad \text{i.e.} \quad y = f(p)\, x + g(p) \quad (p = y') \quad (50)$$

is a little more complicated than (47). The equation is linear in the variables x and y but it is non-linear in the sense of the definition given in Sec. 4. After the equation has been differentiated we get

$$dy = p\, dx = f'(p)\, dp\, x + f(p)\, dx + g'(p)\, dp$$

that is

$$[p - f(p)]\, \frac{dx}{dp} = f'(p)\, x + g'(p)$$

If $f(p) \neq p$ we can divide both sides of the equation by $p - f(p)$ and thus obtain a linear equation [see equation (19)] in which x is regarded as a function of p. After the last equation has been integrated we obtain a relationship of the form $x = x(p, C)$. This relation together with relation (50) defines the general solution of the original equation in parametric form.

In a special case when $f(p) \equiv p$ equation (50) is called **Clairaut's equation** after the French mathematician A. Clairaut (1713-1765)

who was the first to investigate the equation in 1734. It has the form $y = xy' + g(y')$, i.e.

$$y = xp + g(p) \quad (p = y') \tag{51}$$

The differentiation yields $p\,dx = p\,dx + x\,dp + g'(p)\,dp$, and thus we have

$$dp\,[x + g'(p)] = 0 \tag{52}$$

Equating the first factor to zero we obtain, by (51),

$$p = C, \quad \text{i.e.} \quad y = Cx + g(C) \tag{53}$$

This is the general solution of equation (51).

Equating to zero the second factor entering into the left-hand side of (52) we deduce

$$x = -g'(p), \quad y = xp + g(p) = -pg'(p) + g(p) \tag{54}$$

Thus we have obtained one more solution, a singular solution, which

is not contained in the general solution and is represented parametrically. Geometrically, formula (53) defines a family of straight lines (why?). Formula (54) defines the envelope of the family. [Verify the last assertion on the basis of equation (53).]

For instance, the equation $y = xy' - y'^2$ has the general solution

$$y = Cx - C^2 \tag{55}$$

and the singular solution whose graph is the envelope of the family of straight lines (55). To find the envelope we differentiate

Fig. 292

both sides of (55) with respect to C which yields

$$0 = x - 2C$$

Eliminating C from the last two formulas we get $C = \frac{x}{2}$. This results in

$$y = \frac{x}{2}\,x - \left(\frac{x}{2}\right)^2 = \frac{x^2}{4}$$

The corresponding integral curves are depicted in Fig. 292.
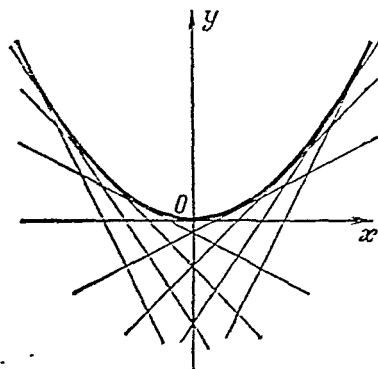
§ *3. Higher-Order Equations and Systems of Differential Equations*

10. Higher-Order Differential Equations. Some general notions related to such equations were given in Sec. 2 [namely, equation (5), general solution (8) or (9) and initial conditions (11)]. By the way, usually it is easier to investigate an nth-order equation when it is given in the form solved for the highest derivative:

$$y^{(n)} = f(x, \, y, \, y', \, \ldots, \, y^{(n-1)})$$

In particular, Cauchy's theorem is immediately extended to such an equation (see Sec. 3): if the function f is continuous and has finite partial derivatives of the first order with respect to y, y', $\ldots$ $\ldots$ $y^{(n-1)}$ at a given initial point defined by conditions (11) then there exists a solution of the equation satisfying the initial conditions and the solution is unique.

We now consider some particular types of higher-order equations that are integrable by quadratures. For higher-order equations such cases are still rarer than for the first-order equations. Here we shall consider non-linear equations (linear equations will be treated in detail in § 4). The main method for formal integration of non-linear higher-order differential equations is the method of reducing the order. It enables us to pass to an equation of a lower order which is equivalent to the original equation. As a rule, the lower the order of an equation, the easier the integration. Besides, we can sometimes come to a first-order equation belonging to one of the integrable types (see Sec. 4) after the order of the original equation is reduced several times. In such a case we are able to complete the integration. Here we shall consider certain particular methods of reducing the order. Some other techniques can be found in [24].

1. For example, let us take the equation of the second order

$$y'^2 + yy'' = 0$$

To integrate it observe that its left-hand side can be rewritten in the form

$$y'^2 + yy'' \equiv (yy')'$$

whence $(yy')' = 0$, $yy' = C_1$, $y \, dy = C_1 \, dx$ and

$$\frac{y^2}{2} = C_1 x + C_2$$

(which is the general solution).

Further, an equation of the form

$$y'^2 - yy'' = 0$$

can be easily integrated in a similar way if we divide both sides by y^2 beforehand. Indeed,

$$\frac{y'^2 - yy''}{y^2} = 0, \quad -\left(\frac{y'}{y}\right)' = 0, \quad \frac{y'}{y} = C_1 \quad \text{and} \quad \frac{dy}{y} = C_1\,dx$$

which results in the general solution

$$\ln|y| = C_1 x + \ln C_2, \quad y = C_2 e^{C_1 x}$$

Here, after the division by y^2, we have obtained a so-called *integrable combination*, that is an expression whose "exact derivative" is equated to zero. Such a method is sometimes applicable to other equations.

For the sake of simplicity, we shall further consider the cases when the order can be reduced for an equation of the second order of the form

$$F(x, y, y', y'') = 0 \tag{56}$$

2. Let an equation of form (56) not contain y. Let it contain only the derivatives of y and the argument. This is an equation of the type

$$F(x, y', y'') = 0 \tag{57}$$

Introducing the notation $y' = p = p(x)$ we obtain the equation

$$F(x, p, p') = 0$$

which is implied by (57). Thus, we have got a first-order equation. Suppose we have managed to integrate this equation and have obtained its general solution $p = \varphi(x, C_1)$. Then we have $y' = \varphi(x, C_1)$ and therefore the general solution of equation (57) is thus obtained:

$$y = \int \varphi(x, C_1)\,dx + C_2$$

3. Let equation (56) not contain x, i.e. let it be of the form

$$F(y, y', y'') = 0 \tag{58}$$

Then we again put $y' = p$ but regard p as being a function of y. It should be noted that here we cannot simply substitute the expression $y'' = p'$ for the derivative y'' into (58) because p' is the derivative of p with respect to x but not to the new argument y. Therefore we write

$$y'' = \frac{d(y')}{dx} = \frac{dp}{dx} = \frac{dp}{dy}\,\frac{dy}{dx} = p\,\frac{dp}{dy}$$

Now we deduce from equation (58) the equation

$$F\left(y, p, p\,\frac{dp}{dy}\right) = 0$$

which is a first-order equation. If we manage to integrate it and to find its general solution $p = \varphi\,(y,\ C_1)$ then we have $\frac{dy}{dx} = \varphi\,(y,\ C_1)$ and therefore the general solution of equation (58) can be directly written in the form

$$\int \frac{dy}{\varphi\,(y,\ C_1)} = x + C_2$$

4. Let the left-hand side of equation (56) be a homogeneous function in the variables y. y' and y'' (see Sec. IX.12):

$$F\,(x,\ ty,\ ty',\ ty'') \equiv t^h F\,(x,\ y,\ y',\ y'') \tag{59}$$

In this case we can reduce the order by means of the substitution

$$\frac{y'}{y} = u = u\,(x)$$

Indeed, the substitution results in

$$y' = uy, \quad y'' = u'y + uy' = u'y + u \cdot uy = (u' + u^2)\,y,$$
$$F\,(x,\ y,\ yu.\ y\,(u' + u^2)) = 0$$

and thus $F\,(x,\ 1,\ u,\ u' + u^2) = 0$. While writing the last expression we have used property (59). If we manage to integrate the first-order equation thus obtained we have

$$u = \varphi\,(x,\ C_1), \quad \frac{y'}{y} = \varphi\,(x,\ C_1)$$

From this we deduce the general solution of the original equation:

$$\ln|y| = \int \varphi\,(x,\ C_1)\,dx + \ln C_2, \text{ i.e. } y = C_2 e^{\int \varphi\,(x,\ C_1)\,dx}$$

11. Connection Between Higher-Order Equations and Systems of First-Order Equations. A higher-order equation of form (5) can always be reduced to a system of n equations of the first order containing n unknown functions. To do this it is sufficient to introduce the notation

$$y = y_1, \quad y' = y_2, \ \ldots, \quad y^{(n-1)} = y_n \tag{60}$$

Then, by equation (5), we can write

$$\left.\begin{array}{l} y_1' = y_2 \\ y_2' = y_3 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ y_{n-1}' = y_n \\ F\,(x,\ y_1,\ y_2,\ \ldots,\ y_n,\ y_n') = 0 \end{array}\right\} \tag{61}$$

System (61) is of a particular form. The general form of a first-order system (for the case of three equations containing three un-

known functions) is written as

$$\left.\begin{array}{l} F_1\,(x,\ y_1,\ y_2,\ y_3,\ y_1',\ y_2',\ y_3') = 0 \\ F_2\,(x,\ y_1,\ y_2,\ y_3,\ y_1',\ y_2',\ y_3') = 0 \\ F_3\,(x,\ y_1,\ y_2,\ y_3,\ y_1',\ y_2',\ y_3') = 0 \end{array}\right\} \qquad (62)$$

The general system of n equations in n unknown functions has a similar form.

Conversely, a system of form (62) can be transformed to one equation (in this case to a third-order equation and in the general case to an nth-order equation) in one unknown function (for example, the function y_1). Therefore the general solution of system (62) contains three arbitrary constants:

$$y_1 = \varphi_1\,(x,\ C_1,\ C_2,\ C_3), \quad y_2 = \varphi_2\,(x,\ C_1,\ C_2,\ C_3),$$
$$y_3 = \varphi_3\,(x,\ C_1,\ C_2,\ C_3)$$

In order to deduce an equation of form (5) (with $n = 3$) from system (62) we should differentiate each of the equations twice. Then together with equalities (62) we obtain nine relations from which we can eliminate the eight quantities y_2, y_3, y_2', y_3', y_2'', y_3'', y_2''', and y_3'''. This yields the sought-for third-order equation. After the equation has been integrated we obtain the general solution $y_1 = \varphi_1\,(x.\ C_1,\ C_2,\ C_3)$. To complete the process of solving the system we must find y_2 and y_3. But this can be performed without integration since y_2 and y_3 are expressed in terms of y_1 and its derivatives by means of the above-mentioned relations.

A general system of differential equations of arbitrary orders can be transformed into a system of first-order equations by introducing notation similar to (60). For example, suppose we have a system of two equations. Let the equations be of the third order with respect to one of the unknown functions and of the second order with respect to the other. Then such a system is equivalent to a system of five first-order equations containing five unknown functions. The general solution of the original equation in this example must contain five arbitrary constants.

12. Geometric Interpretation of System of First-Order Equations. For simplicity's sake, let us take the case of a system of two first-order equations in two unknown functions $y_1\,(x)$ and $y_2\,(x)$:

$$\left.\begin{array}{l} F_1\,(x,\ y_1,\ y_2,\ y_1',\ y_2') = 0 \\ F_2\,(x,\ y_1,\ y_2,\ y_1',\ y_2') = 0 \end{array}\right\} \qquad (63)$$

If it is possible to solve the system for y_1' and y_2' before performing integration we obtain a system of the form

$$\left.\begin{array}{l} y_1' = f_1\,(x,\ y_1,\ y_2) \\ y_2' = f_2\,(x,\ y_1,\ y_2) \end{array}\right\} \qquad (64)$$

Then we say that the system is written in the **normal form**.

A solution of (63), or of (64), is, of course, a pair of functions

$$y_1 = y_1(x) \quad \text{and} \quad y_2 = y_2(x) \tag{65}$$

which converts both equations into identities. According to Sec. 11, the general solution of the system contains two arbitrary constants, that is it has the form

$$y_1 = y_1(x, C_1, C_2), \quad y_2 = y_2(x, C_1, C_2)$$

System of equations (64) and its solution (65) can be simply interpreted geometrically if we introduce a three-dimensional Cartesian space with the coordinates x, y_1 and y_2 and regard it as a usual geometric space. Then formulas (65) yield the parametric representation of a curve (see Sec. VII.23) if we consider x to be a parameter (we can formally write an additional equation of the form $x = x$). This curve is called an integral curve of system of equations (64). Let us take an arbitrary point M $(x,\ y_1,\ y_2)$ in the space (see Fig. 293) and calculate the values of the right-hand sides of system (64) at the point. These values being equal to y_1' and y_2', we thus determine the directions of the tangent lines to the curves $y_1 = y_1(x)$

Fig. 293

(L) is an integral curve

and $y_2 = y_2(x)$, i.e. to the projections of the integral curve. Therefore we can determine the direction of the tangent to the integral curve at the point M provided it passes through M. Point M being an arbitrary point, we see that system (64) defines a direction field in the x, y_1, y_2-space. Hence, an integral curve is a curve whose tangent goes along the direction of the field at each point (compare with Sec. 3).
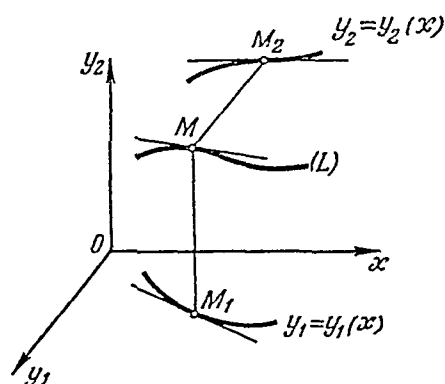
The variables y_1 and y_2 are involved equivalently in system (64) whereas the variable x plays a specific role. But there are such cases when all the three variables are equivalent so that each of them can be taken as an independent variable. It is preferable to write a system of this kind in the so-called symmetric form

$$\frac{dx}{P(x, y, z)} = \frac{dy}{Q(x, y, z)} = \frac{dz}{R(x, y, z)} \tag{66}$$

It is easy to obtain form (64) from form (66) and vice versa (how can we do this?).

The geometric interpretation of system (66) is analogous to the above interpretation of system (64). By relations (66), the vector $dr = dx\mathbf{i} + dy\mathbf{j} + dz\mathbf{k}$ must be parallel to the known vector $P\mathbf{i} + Q\mathbf{j} + R\mathbf{k}$ at each given point $M(x, y, z)$ (why?). The problem of integrating system (66) is therefore reduced to the problem of constructing curves in space which have prescribed directions of their tangents at each point.

The geometric meaning of system (64) implies that in order to isolate a unique integral curve we must specify a point $M_0(x_0, y_0, z_0)$ in space through which the curve should pass. In other words, the initial condition

$$y_1(x_0) = y_{1,0}, \quad y_2(x_0) = y_{2,0}$$

uniquely defines a certain solution of system (64). As in Sec. 7, in the case of a system of equations we can also encounter singular points and singular curves. These points and curves can be distinguished in a manner similar to the one described in Sec. 7. In particular, a *singular point of system* (66) is a point at which all the three denominators vanish. Every point of this kind is a singular point; the vector $P\mathbf{i} + Q\mathbf{j} + R\mathbf{k}$ turns into zero at such a point and thus it has no certain direction at the point.

A normal system of first-order equations with an arbitrary number of unknown functions has the general form

$$\left.\begin{array}{l} y_1' = f_1(x, y_1, y_2, \ldots, y_n) \\ y_2' = f_2(x, y_1, y_2, \ldots, y_n) \\ \cdots \cdots \cdots \cdots \cdots \cdots \cdots \\ y_n' = f_n(x, y_1, y_2, \ldots, y_n) \end{array}\right\} \quad (67)$$

A solution of the system is a set of functions of the form

$$y_1 = y_1(x), \quad y_2 = y_2(x), \ldots, \quad y_n = y_n(x) \quad (68)$$

which reduces the system to an identity. The general solution of such a system contains n arbitrary constants.

To specify a unique solution we can set initial conditions of the form

$$y_1(x_0) = y_{1,0}, \quad y_2(x_0) = y_{2,0}, \ldots, \quad y_n(x_0) = y_{n,0} \quad (69)$$

Cauchy proved that there is a solution of system (67) satisfying conditions (69) and that the solution is unique if the right-hand sides of system (67) are continuous and possess finite partial derivatives of the first order with respect to the variables $y_1, y_2, \ldots, y_n$ for the values $x = x_0, y_1 = y_{1,0}, \ldots, y_n = y_{n,0}$.

The geometric meaning of system (67), its solution (68) and conditions (69) is that the system defines a direction field in an $(n+1)$-dimensional space with the coordinates $x, y_1, \ldots, y_n$ and that

the solution of the system satisfying conditions (69) defines an integral curve lying in the space and passing through the point $M_0 (x_0, y_{1,0}, \ldots, y_{n,0})$ determined by (69) (see Sec. VII.18).

In case the right-hand sides of system (67) do not depend on the variable x the system is called **autonomous**. It turns out that it is more convenient to regard its solutions as curves lying in the n-dimensional space of the variables $y_1, y_2, \ldots, y_n$ which is a subspace of the $x, y_1, \ldots, y_n$-space, the variable x playing the role of a parameter. In such a case the space of variables $y_1, \ldots, y_n$ is called the **phase space** of the system and the curve $y_1 (x), \ldots \ldots, y_n (x)$ is called a **phase trajectory** of the system. A point $(y_1, \ldots, y_n)$ of the n-dimensional space is called a state of the system and thus the space $y_1, \ldots, y_n$ may be called a state space. For simplicity's sake, let us limit our attention to the case $n = 2$. We shall denote the independent variable by t and interpret it as time. The unknown functions y_1 and y_2 will be designated by the letters x and y, i.e. $x = x (t)$ and $y = y (t)$. Then, in place of system (67), we get a system of the form

$$\frac{dx}{dt} = P (x, y), \qquad \frac{dy}{dt} = Q (x, y)$$

Multiplying the first equation by **i** and the second equation by **j** and adding together (in the vectorial sense) the results we arrive at the vector differential equation

$$\frac{d\mathbf{r}}{dt} = \mathbf{A} (x, y), \qquad \text{or} \qquad \frac{d\mathbf{r}}{dt} = \mathbf{A} (\mathbf{r}) \tag{70}$$

where $\mathbf{A} = P (x, y) \mathbf{i} + Q (x, y) \mathbf{j}$ is a given vector field in the phase plane x, y. The derivative $\frac{d\mathbf{r}}{dt}$ being the velocity (see Sec. VII.23), we see that there is a **velocity field** defined in the x, y-plane. Each solution $\mathbf{r} (t) = x (t) \mathbf{i} + y (t) \mathbf{j}$ defines the law of motion of a point in the plane, and the point has a prescribed velocity at each of its positions. We can imagine that equation (70) defines a flow of liquid in the phase plane and that to solutions of the equation there correspond laws of motion of particles of the liquid (but in fact the equations of motion of a liquid medium are much more complicated; such equations are deduced in hydrodynamics and they prove to be partial differential equations). The autonomy of equation (70) implies that the "flow" is stationary and therefore any two distinct trajectories have no common points.

For example, let us write equation (4) in the form of an autonomous system of first-order equations:

$$\frac{dy}{dt} = v, \qquad M \frac{dv}{dt} = -ky \tag{71}$$

where y is the coordinate of the oscillating point and v is its velocity. In courses on physics one can find the following formula of the

total energy of an oscillating point:

$$E = \frac{Mv^2}{2} + \frac{ky^2}{2} \tag{72}$$

(let the reader think about the structure of the formula). In the process of free oscillations without friction the energy must be preserved. Indeed, computing the derivative of E with respect to t on the basis of formula (71) we obtain

$$\frac{dE}{dt} = Mv\frac{dv}{dt} + ky\frac{dy}{dt} = -kyv + kyv = 0$$

This is a mathematical proof of the law of conservation of energy for our example. Hence, we have $E = \mathrm{const}$ for every solution of system (71). Therefore we see that a point representing the state of the system in the phase plane describes an ellipse in the y, v-plane. Different ellipses correspond to different possible oscillations of the material point, and to different ellipses there correspond oscillations about the equilibrium position with different amplitudes (see Fig. 294).
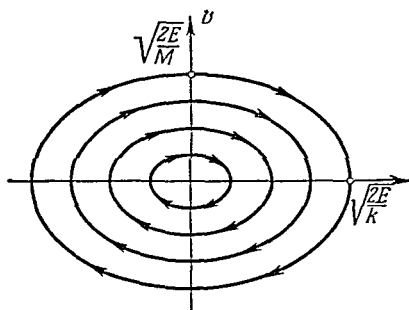
Fig. 294

13. **First Integrals.** For definiteness, we now consider a system of three first-order equations of form (62). Every relation of the form

$$\Phi(x, y_1, y_2, y_3, C) = 0 \quad (\Phi \not\equiv 0) \tag{73}$$

which is identically satisfied by any solution of the system is called a first integral of the system of equations. Here C is a constant which, in general, is different for different solutions.

Knowing a first integral we can reduce the number of equations in the system by unity. Indeed, if we express y_3 in terms of the rest by means of relation (73) and then substitute the result into the first two equations (62) we obtain a system of two first-order equations containing two unknown functions y_1 and y_2. If we integrate the last system, i.e. if we find $y_1(x)$ and $y_2(x)$, then $y_3(x)$ can be found without integration on the basis of equality (73).

Similarly, if we know two independent first integrals we can reduce the number of equations by two. Finally, if we manage to find three independent first integrals (that is such integrals that

none of them is a consequence of the others) of the form

$$\left.\begin{array}{l} \Phi_1\,(x,\ y_1,\ y_2,\ y_3,\ C_1) = 0 \\ \Phi_2\,(x,\ y_1,\ y_2,\ y_3,\ C_2) = 0 \\ \Phi_3\,(x,\ y_1,\ y_2,\ y_3,\ C_3) = 0 \end{array}\right\}$$

then we obtain the general solution of system (62) represented in an implicit form.

It is sometimes possible to find first integrals by deducing so-called **integrable combinations** from given equations. For instance, we can easily derive such a combination for the system

$$\left.\begin{array}{l} y' = y + z \\ z' = -y + z \end{array}\right\} \tag{74}$$

Indeed, we see that

$$yy' + zz' = y\,(y + z) + z\,(-y + z) = y^2 + z^2$$

and therefore

$$\frac{1}{2}\,(y^2 + z^2)' = y^2 + z^2, \qquad \frac{d\,(y^2 + z^2)}{y^2 + z^2} = 2dx \ \text{ and}$$

$$\ln\,(y^2 + z^2) = 2x + \ln C$$

Finally, we obtain the first integral

$$y^2 + z^2 = Ce^{2x}$$

It follows that if $x \to \infty$ then every solution approaches infinity, and if $x \to -\infty$ it tends to zero. As a rule, in other cases we can also draw some important conclusions concerning the behaviour of solutions without completing integration of a system in question when we know one or more first integrals of the system. Returning to system (74) we see that we can receive another first integral if we divide one of the equations by the other (let the reader perform the calculations!).

There are some cases when first integrals can be derived on the basis of certain physical considerations, more often by different conservation laws.

For instance, relation (72) is a first integral of system (71) because we can regard E as an arbitrary constant C. Expressing v in terms of y from the relation and substituting the result into the first equation (71) we can easily complete the integration (let the reader do this!).

In conclusion, we note that, as it has been seen, it is more natural to consider systems of equations in which the number of equations coincides with the number of unknown functions. Such systems are called closed. If the number of unknown functions exceeds the number of equations the system is called non-closed (or sub-definite). The excessive number of unknown functions can be chosen arbi-

trarily. The fact that a system is non-closed usually indicates that some necessary relations were not taken into account. If the number of equations is greater than the number of unknown functions the system is called overdetermined. Such a system is usually contradictory, that is it has no solutions. An overdetermined system can also indicate that there is an interdependence between the equations entering into the system in question. This means that some of the equations are consequences of the rest and therefore they are unnecessary and can be dropped. Such a situation may also show that when deducing the equations we made a mistake.

§ 4. Linear Equations of General Form

14. Homogeneous Linear Equations. The methods of investigating linear equations of arbitrary orders have many features similar to those of the methods of solving first-order linear equations (see Sec. 4). But in the general case it is no longer possible to integrate an equation by quadratures. For the sake of simplicity, we first consider second-order equations. The equation

$$z'' + p(x) z' + q(x) z = 0 \qquad (75)$$

whose left-hand side is a linear function in the unknown function and its derivatives is called a homogeneous linear equation.

For brevity, let us denote the left-hand side of equation (75) as $L[z]$, i.e. in this case, by definition,

$$L[z] \equiv z'' + p(x) z' + q(x) z$$

Then equation (75) can be rewritten in the form

$$L[z] = 0$$

The expression $L[z]$ possesses the following properties:

$$L[z_1 + z_2] = (z_1 + z_2)'' + p(x)(z_1 + z_2)' + q(x)(z_1 + z_2) =$$
$$= (z_1'' + p(x) z_1' + q(x) z_1) + (z_2'' + p(x) z_2' + q(x) z_2) =$$
$$= L[z_1] + L[z_2], \qquad L[Cz] = CL[z]$$

where $C = \text{const}$ (the last property is verified similarly).

Expressions of this type are called linear operators; we mentioned them in Sec. XIV.26.

We can easily prove the following properties of equation (75):

1. A sum of solutions of equation (75) is a solution of the same equation. Actually, if z_1 and z_2 are two solutions of equation (75) then

$$L[z_1] = 0 \quad \text{and} \quad L[z_2] = 0$$

and thus

$$L[z_1 + z_2] = L[z_1] + L[z_2] = 0$$

2. If we multiply a solution of equation (75) by a constant we get a solution of the same equation. The verification of the second property is analogous to that of the first property.

Now we can combine properties 1 and 2 in the following way: a linear combination of solutions of equation (75) (see Sec. VII.5) is a solution of the same equation. For instance, if $z_1(x)$ and $z_2(x)$ satisfy equation (75) then

$$z = C_1 z_1(x) + C_2 z_2(x) \tag{76}$$

also satisfies the same equation for any values of the constants C_1 and C_2.

3. A function which is identically equal to zero always satisfies equation (75).

4. If we know a nonzero solution of equation (75) we can reduce by unity the order of the equation. Virtually, let $z_1(x)$ be such a solution. Make a substitute of the form $z = z_1 u$ where $u = u(x)$ is a new unknown function. Then we get

$$(z_1'' u + 2 z_1' u' + z_1 u'') + p(z_1' u + z_1 u') + q z_1 u = 0$$

that is

$$z_1 u'' + (2 z_1' + p z_1) u' + (z_1'' + p z_1' + q z_1) u = 0$$

But since $L[z_1] = 0$, the last term vanishes and hence, after the substitution $u' = v$, we finally receive

$$z_1 v' + (2 z_1' + p z_1) v = 0$$

which is a first-order linear equation, that is an equation whose order is less by unity than the order of the original equation.

Now we complete the integration:

$$\frac{dv}{v} = -\frac{2 z_1' + p z_1}{z_1} dx, \quad \ln|v| = -2 \ln|z_1| - \int p(x)\,dx + \ln C_2,$$

$$v = \frac{C_2}{z_1^2} e^{-\int p(x)dx}, \quad u = C_2 \int \frac{1}{z_1^2} e^{-\int p(x)dx}\,dx + C_1$$

and hence

$$z = C_1 z_1 + C_2 z_1 \int \frac{1}{z_1^2} e^{-\int p(x)dx}\,dx \tag{77}$$

The function in front of which the factor C_2 is placed is one of the particular solutions of equation (75) because we can obtain it from general solution (77) by putting $C_1 = 0$ and $C_2 = 1$. Therefore, denoting this solution by z_2 we arrive at the fifth property.

5. The general solution of equation (75) has form (76) where C_1 and C_2 are arbitrary constants and z_1 and z_2 are two particular solutions of the equation.

It should be noted that formula (76) in fact expresses the general solution of equation (75) if and only if the solutions z_1 and z_2 are linearly independent. The concept of linear dependence of functions is similar to that of vectors (see Sec. VII.5). Namely, we say that several given functions are linearly dependent if one of them is a linear combination of the rest. In particular, two functions $z_1(x)$ and $z_2(x)$ are linearly dependent if and only if $z_2(x) \equiv Cz_1(x)$ or $z_1(x) \equiv Cz_2(x)$. Thus, $z_1(x)$ and $z_2(x)$ are proportional to each other in this case. Hence, we see that formula (76) does not yield the general solution in such a case because

$$C_1 z_1 + C_2 z_2 \equiv C_1 z_1 + C_2 C z_1 \equiv (C_1 + C_2 C) z_1(x) = D z_1(x)$$

where $D = C_1 + C_2 C$ is a constant. This means that although formally we have two arbitrary constants on the right-hand side of expression (76), these constants are not essential parameters since their number can be reduced by unity (see Sec. X.2).

All the enumerated properties also hold for a homogeneous linear equation of any order of the form

$$z^{(n)} + p(x) z^{(n-1)} + q(x) z^{(n-2)} + \ldots + s(x) z = 0 \qquad (78)$$

with the only exception of the fifth property which must be replaced by the following assertion: the general solution of equation (78) has the form

$$z = C_1 z_1(x) + C_2 z_2(x) + \ldots + C_n z_n(x) \qquad (79)$$

instead of (76). The functions $z_1(x)$, $z_2(x)$, $\ldots$, $z_n(x)$ entering into formula (79) are any n linearly independent solutions of equation (78) and C_1, C_2, $\ldots$, C_n are arbitrary constants. Any family of n linearly independent solutions of equation (78) is called its **fundamental system of solutions.** Thus, *the general solution of equation (78) is a linear combination of solutions forming a fundamental system of solutions with n arbitrary constant coefficients.* Using the terminology of Secs. VII.17-19 we can say that the totality of all the solutions of equation (78) is an n-dimensional linear space. A fundamental system of solutions is a basis in such a space.

In conclusion we note that the left-hand side of equation (78) is a homogeneous function in the variables $z, \ldots, z^{(n-1)}, z^{(n)}$ and therefore we can reduce by unity the order of the equation with the help of the method of Sec. 10 (see type 4). But this procedure is rarely performed because after the order is reduced the equation becomes non-linear.

15. Non-Homogeneous Equations. We now consider a non-homogeneous linear equation of the form

$$y'' + p(x) y' + q(x) y = f(x) \qquad (80)$$

According to Sec. 14, let us denote the left-hand side by $L[y]$.

1. A particular solution of equation (80) being known, it is possible to reduce the problem of integrating the equation to the problem of integrating a homogeneous equation (75) which corresponds to (80) [that is an equation of form (75) having the same coefficients on the left-hand side as equation (80) but with zero right-hand side, or, simply, an equation that is obtained from (80) by dropping its right-hand side].

Indeed, if $Y(x)$ is such a solution then after the substitution

$$y = Y(x) + z \tag{81}$$

where $z = z(x)$ is a new unknown function we obtain

$$L[Y + z] = f(x)$$

which suggests

$$L[Y] + L[z] = f(x)$$

But $L[Y] = f(x)$ (why is it so?). Hence we arrive at an equation of form (75) containing z as an unknown function.

Thus, *the general solution of non-homogeneous linear equation* (80) *is the sum of any particular solution of the equation and the general solution of the corresponding homogeneous linear equation* (compare this result with the general solution of the first-order equation considered in Sec. 4).

2. If the right-hand side $f(x)$ is a linear combination of some functions, for instance, of two functions, that is $f(x) = \alpha f_1(x) + \beta f_2(x)$ (where α and β are constants), and if certain particular solutions $Y_1(x)$ and $Y_2(x)$ of equation (80) having right-hand sides equal to $f_1(x)$ and $f_2(x)$, respectively, are known, then the function

$$Y(x) = \alpha Y_1(x) + \beta Y_2(x)$$

is a particular solution of equation (80) with the right-hand side $f(x)$.

This simple property is in fact a special case of a general principle called the superposition principle (see Sec. XIV.26). The proof of the property is left to the reader.

3. If the general solution of homogeneous equation (75) is known the general solution of equation (80) can be found by quadratures.

This is performed by means of a method discovered by Lagrange. It is called the *method of variation of parameters* (we used the method in Sec. 4 in solving the first-order non-homogeneous linear equation). As we know, the general solution of equation (75) is of form (76). By analogy with formula (22), we seek a solution of equation (80) in the form

$$y = \varphi_1(x) z_1(x) + \varphi_2(x) z_2(x) \tag{82}$$

where φ_1 and φ_2 are some functions yet unknown. But there are two such functions here and only one equation (80). Therefore, in order to find the functions, we shall impose one more additional condition which will be put down below [condition (84)].

Differentiating equality (82) we obtain

$$y' = (\varphi_1 z_1' + \varphi_2 z_2') + (\varphi_1' z_1 + \varphi_2' z_2) \tag{83}$$

We now set the condition that the expression inside the second parentheses on the right-hand side of relation (83) should identically vanish:

$$\varphi_1' z_1 + \varphi_2' z_2 = 0 \tag{84}$$

Then when differentiating equality (83), we must take into account only the expression entering into the first parentheses, that is

$$y'' = (\varphi_1 z_1'' + \varphi_2 z_2'') + (\varphi_1' z_1' + \varphi_2' z_2') \tag{85}$$

Substituting the results (82), (83) and (85) thus obtained into equation (80) and dropping the sum which equals zero we receive

$$\varphi_1 \, (z_1'' + p z_1' + q z_1) + \varphi_2 \, (z_2'' + p z_2' + q z_2) + \\ + (\varphi_1' z_1' + \varphi_2' z_2') = f \, (x)$$

(check up the calculations!).

The functions z_1 and z_2 satisfying equation (75), the first two parentheses in the last equation can be deleted. Thus we arrive at the equality

$$\varphi_1' z_1' + \varphi_2' z_2' = f \, (x) \tag{86}$$

Hence, now we have two relations (84) and (86) for determining φ_1 and φ_2. The functions z_1, z_2 and $f \, (x)$ being regarded as known, we have a system of two algebraic equations of the first degree in two unknown quantities, i.e. in the derivatives $\varphi_1' \, (x)$ and $\varphi_2' \, (x)$. Solving the system we find the unknowns and then, integrating, we obtain φ_1 and φ_2.

For instance, let us take the simplest equation of *forced oscillations* which is obtained if we add an external force $P \, (t)$ to the right-hand side of equation (3). Dividing both sides of the equation by M we derive

$$y'' + \omega_0^2 y = f \, (t) \tag{87}$$

where the notation $\omega_0^2 = \dfrac{k}{M}$ and $f \, (t) = \dfrac{P \, (t)}{M}$ is introduced.

The corresponding homogeneous equation

$$z'' + \omega_0^2 z = 0 \tag{88}$$

has two linearly independent particular solutions of the form $z_1 = \cos \omega_0 t$ and $z_2 = \sin \omega_0 t$ (this fact can be directly verified by substituting the solutions into the equation). The general solution can therefore be written as

$$z = C_1 \cos \omega_0 t + C_2 \sin \omega_0 t \tag{89}$$

In particular, it follows that ω_0 is the **fundamental frequency** (**natural frequency**) of the system in question, that is the frequency of free oscillations arising in the system when there are no external forces.

According to formula (82), we seek a solution of equation (87) in the form

$$y = \varphi_1 (t) \cos \omega_0 t + \varphi_2 (t) \sin \omega_0 t \tag{90}$$

Then equations (84) and (86) turn into

$$\left. \begin{array}{l} \varphi_1' \cos \omega_0 t + \varphi_2' \sin \omega_0 t = 0 \\ \varphi_1' (-\omega_0 \sin \omega_0 t) + \varphi_2' \omega_0 \cos \omega_0 t = f(t) \end{array} \right\}$$

From this we immediately find

$$\varphi_1'(t) = -\frac{1}{\omega_0} f(t) \sin \omega_0 t \quad \text{and} \quad \varphi_2'(t) = \frac{1}{\omega_0} f(t) \cos \omega_0 t$$

Now we must integrate the last relations. It is inconvenient to use indefinite integrals here because they contain arbitrary constants which are difficult to specify. It is therefore better to put down the results in the form of definite integrals with a fixed lower limit of integration and with a variable upper limit. Let the lower limit be equal to zero (we assume that $t = 0$ is the initial instant of time). Then we have

$$\varphi_1 (t) = -\frac{1}{\omega_0} \int_0^t f(t) \sin \omega_0 t \, dt + C_1$$

where C_1 is an arbitrary constant. Here the variable t is understood in two-fold sense because the letter t designates the variable of integration and the upper limit as well. It is therefore more convenient to use the fact that the value of a definite integral is independent of the notation of the variable of integration (see Sec. XIV.3) and rewrite the expression of $\varphi_1 (t)$ in the form

$$\varphi_1 (t) = -\frac{1}{\omega_0} \int_0^t f(\tau) \sin \omega_0 \tau \, d\tau + C_1$$

The expression of $\varphi_2 (t)$ is found similarly:

$$\varphi_2 (t) = \frac{1}{\omega_0} \int_0^t f(\tau) \cos \omega_0 \tau \, d\tau + C_2$$

Substituting φ_1 and φ_2 thus found into (90) we derive

$$y = -\frac{1}{\omega_0} \cos \omega_0 t \int\limits_0^t f(\tau) \sin \omega_0 \tau \, d\tau +$$

$$+\frac{1}{\omega_0} \sin \omega_0 t \int\limits_0^t f(\tau) \cos \omega_0 \tau \, d\tau + C_1 \cos \omega_0 t + C_2 \sin \omega_0 t$$

We now insert $\cos \omega_0 t$ and $\sin \omega_0 t$ under the integral sign. We could not do this if we had not changed the notation of the variable of integration. But now this is permissible and thus we get the expression

$$y = \frac{1}{\omega_0} \int\limits_0^t [-f(\tau) \cos \omega_0 t \sin \omega_0 \tau +$$

$$+ f(\tau) \sin \omega_0 t \cos \omega_0 \tau] \, d\tau + C_1 \cos \omega_0 t + C_2 \sin \omega_0 t$$

in which both integrals are combined. From this we obtain the general solution of equation (87):

$$y = \frac{1}{\omega_0} \int\limits_0^t \sin \omega_0 (t - \tau) f(\tau) \, d\tau + C_1 \cos \omega_0 t + C_2 \sin \omega_0 t \qquad (91)$$

The arbitrary constants C_1 and C_2 can be determined if we set, for example, the initial conditions

$$y(0) = y_0, \quad y'(0) = v_0 \qquad (92)$$

Substituting $t = 0$ into both sides of (91) we obtain $y_0 = C_1$. To use the second condition (92) we must differentiate equality (91) with respect to t and then substitute $t = 0$. When differentiating the integral we must take into account that t enters into the integral as an upper limit of integration and as a parameter under the integral sign. Hence, using formula (XIV.80). we deduce

$$y' = \frac{1}{\omega_0} \int\limits_0^t \omega_0 \cos \omega_0 (t - \tau) f(\tau) \, d\tau +$$

$$+ \left[\frac{1}{\omega_0} \sin \omega_0 (t - \tau) f(\tau) \right]_{\tau=t} - C_1 \omega_0 \sin \omega_0 t + C_2 \omega_0 \cos \omega_0 t =$$

$$= \int\limits_0^t \cos \omega_0 (t - \tau) f(\tau) \, d\tau - C_1 \omega_0 \sin \omega_0 t + C_2 \omega_0 \cos \omega_0 t$$

Thus we obtain

$$v_0 = C_2 \omega_0, \quad \text{i.e.} \quad C_2 = \frac{v_0}{\omega_0}$$

Consequently, the particular solution of equation (87) satisfying the initial conditions (92) is

$$y = \frac{1}{\omega_0} \int\limits_0^t \sin \omega_0 (t - \tau) f (\tau) \, d\tau + y_0 \cos \omega_0 t + \frac{v_0}{\omega_0} \sin \omega_0 t$$

All the enumerated properties hold for the general equation

$$y^{(n)} + p(x) y^{(n-1)} + q(x) y^{(n-2)} + \ldots + s(x) y = f(x) \qquad (93)$$

as well. The method of variation of arbitrary constants is applied to equation (93) in the following way: we substitute functions $\varphi_1(x)$, $\varphi_2(x)$, ..., $\varphi_n(x)$ for C_1, C_2, ..., C_n, respectively, into formula (79) and then differentiate the formula in succession up to the $(n - 1)$th derivative inclusive. After each of the differentiations we obtain a group of terms containing the derivatives φ_1', φ_2', ..., φ_n', and we equate each of the groups to zero. Finally, differentiating formula (79) [with C_1, C_2, ..., C_n replaced by $\varphi_1(x)$, $\varphi_2(x)$, ..., $\varphi_n(x)$] the nth time and substituting all the expressions thus obtained into equation (93) we derive the nth relation connecting φ_1', φ_2', ..., φ_n'. Then φ_1', φ_2', ..., φ_n' are found by solving the linear algebraic system of equations etc.

4. Any solution of equation (78) or of equation (93) can be extended to any interval in which the coefficients and the right-hand side of the equation do not approach infinity. Generally, this is not the case for non-linear equations because it can happen that a solution (or its derivatives) approaches infinity for some finite value of x. A simple example of this fact is equation (38) in the case $\alpha > 1$ (see Fig. 288). Here the slope of the direction field increases so fast, as y increases, that integral curves travel into infinity after passing only a finite distance along the x-axis. On the contrary, a solution of a linear equation (for instance, the solution $y = Ce^{Mx}$ of the equation $y' = My$) cannot approach infinity for a finite value of x.

16. Boundary-Value Problems. In the preceding section we studied problems in which we isolated a particular solution from the general solution by means of initial conditions which define certain values of an unknown function and of its derivatives for a single value of the argument. But there are some other ways of isolating a particular solution from the general solution which are encountered in practical problems. At the same time it is common for all methods of determining a particular solution that the number of additional conditions imposed on a sought-for solution must be equal to the number of degrees of freedom (see Sec. X.2) in the general solution of an equation in question, that is to the order of the equation.

These additional conditions can be written, in the case of equation (5) of the nth order, in the form

$$G_k[y] = \alpha_k \qquad (k = 1, 2, \ldots, n) \qquad (94)$$

where $G_k[y]$ $(k = 1, 2, \ldots, n)$ is a given combination of the values of the sought-for function $y(x)$ and its derivatives taken for different values of the argument, in the general case, and α_1, $\alpha_2, \ldots, \alpha_n$ are given numbers. [More precisely, $G_k[y]$ $(k = 1, 2, \ldots, n)$ is a given functional, the notion of a functional being mentioned in Sec. XIV.4.] For instance, if we take the case of initial conditions (11) then $G_k[y]$ is simply equal to $y^{(k-1)}(x_0)$.

If general solution (8) of a given equation is known then to find the particular solution we are interested in we must substitute the expression of the general solution into conditions (94) which results in a system of n equations in n unknowns $C_1, C_2, \ldots, C_n$. If

$$G_k[C_1y_1 + C_2y_2] = C_1G_k[y_1] + C_2G_k[y_2] \quad (C_1, C_2 = \text{const})$$

then conditions (94) are called linear. If, in addition, all $\alpha_k = 0$ then the conditions are called homogeneous linear conditions. If we have functions (which may not necessarily be solutions of the differential equation) satisfying homogeneous linear conditions then any linear combination of the functions satisfies the conditions as well. In fact, if, for example, $G_k[y_1] = 0$ and $G_k[y_2] = 0$ then

$$G_k[C_1y_1 + C_2y_2] = C_1G_k[y_1] + C_2G_k[y_2] = C_1 \cdot 0 + C_2 \cdot 0 = 0$$

The difference of two functions satisfying the same non-homogeneous linear conditions satisfies the corresponding homogeneous condition (that is a homogeneous linear condition with the same left-hand side). Let the reader verify the assertion!

In our further discussion we shall limit ourselves to solutions of the equation

$$y'' + p(x) y' + q(x) y = f(x) \qquad (a \leqslant x \leqslant b) \tag{95}$$

with the additional conditions

$$y(a) = \alpha_1, \qquad y(b) = \alpha_2 \tag{96}$$

although all the general conclusions we shall draw remain true for linear equations of any order n with additional linear conditions (94) of arbitrary form. The interval (a, b) will be regarded as being finite. We shall also assume that the functions $p(x)$, $q(x)$ and $f(x)$ are finite. Then we can regard any solution as extended over the whole interval including its end-points (see property 4 in Sec. 15). Conditions of form (96) containing only the end-point values of functions defined over the interval in which the solution is sought for are called **boundary conditions**. The corresponding problem of determining the solution of the equation is called a **boundary-value problem**.

The solution of the above boundary-value problem is found on the basis of the general solution of equation (95) which is

$$y(x) = Y(x) + C_1z_1(x) + C_2z_2(x) \tag{97}$$

where $Y(x)$ is a certain particular solution of equation (95) and z_1 and z_2 are two linearly independent particular solutions of the corresponding homogeneous equation (see property 1 in Sec. 15). Substituting formula (97) into condition (96) we obtain two relations for C_1 and C_2:

$$\left.\begin{array}{l} C_1 z_1(a) + C_2 z_2(a) = \alpha_1 - Y(a) \\ C_1 z_1(b) + C_2 z_2(b) = \alpha_2 - Y(b) \end{array}\right\} \tag{98}$$

In solving this system of two algebraic equations of the first degree we can encounter the following two cases (see Secs. VI.4 and VI.6):

1. *Basic case.* The determinant of the system is unequal to zero. Since in this case system (98) has a certain uniquely defined solution, equation (95) with conditions (96) possesses one and only one solution for any right-hand member $f(x)$ and any numbers α_1, α_2.

2. *Singular case.* The determinant of the system is equal to zero. In such a case system (98) is, as a rule, contradictory but it may have infinitely many solutions for certain specific right-hand sides. Hence, equation (95) with conditions (96) has, as a rule, no solution when the function $f(x)$ and the numbers α_1, α_2 are chosen arbitrarily but it has an infinitude of solutions when $f(x)$, α_1, α_2 are chosen in a certain specific manner. For example, we can easily verify that if $f(x)$ and α_1 are given beforehand then there are infinitely many solutions only for one specific value of α_2, the problem having no solutions for any other value of α_2.

It should be noted that it is the form of the left-hand sides of equation (95) and of conditions (96) that determines which of the above cases takes place.

According to Sec. VI.6, the basic case takes place if and only if the corresponding homogeneous problem [in which we must put $f(x) \equiv 0$, $\alpha_1 = 0$ and $\alpha_2 = 0$] has only the zero solution. In the singular case the homogeneous problem has infinitely many solutions. If the non-homogeneous problem possesses at least one solution then adding the general solution of the corresponding homogeneous problem to the particular solution of the non-homogeneous problem we obtain the general solution of the non-homogeneous problem.

When we deal with an initial-value problem (Cauchy's problem) we always have the basic case because the solution always exists and is unique. In solving a boundary-value problem we can encounter the singular case as well. For instance, let us consider the following problem containing a constant parameter $\lambda = \text{const}$:

$$y'' + \lambda y = f(x) \quad (0 \leqslant x \leqslant l), \quad y(0) = \alpha_1, \quad y(l) = \alpha_2 \tag{99}$$

Let us first take the case $\lambda > 0$. Then the functions $z_1(x) = \cos \sqrt{\lambda} x$ and $z_2(x) = \sin \sqrt{\lambda} x$ are two linearly independent solutions of the

corresponding homogeneous differential equation and hence the determinant of system (98) is equal to

$$\begin{vmatrix} z_1(0) & z_2(0) \\ z_1(l) & z_2(l) \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ \cos\sqrt{\lambda}\,l & \sin\sqrt{\lambda}\,l \end{vmatrix} = \sin\sqrt{\lambda}\,l$$

Equating the determinant to zero we find the values

$$\lambda = \left(\frac{\pi}{l}\right)^2, \quad \left(\frac{2\pi}{l}\right)^2, \quad \left(\frac{3\pi}{l}\right)^2, \ldots \tag{100}$$

for which there is a singular case for problem (99). This means that either the existence or the uniqueness of the solution is violated.

The set of values of a parameter entering into the statement of a problem for which the problem degenerates in a certain sense (see Sec. II.8) is called the spectrum of the problem. We suggest that the reader should verify that in the case $\lambda \leqslant 0$ we always have the basic case for problem (99). This means that the set of the values (100) is the spectrum of the problem.

(a)

(b)

Fig. 295

The result obtained above has an important application to the problem of investigating the stability of an elastic bar when it is subjected to buckling by a central force. Let a homogeneous (i.e. having the same properties in all its parts) elastic weightless bar be placed along the x-axis. Suppose that the bar is compressed by a force P and that its ends can rotate about the points of support which are permanently kept on the x-axis (see Fig. 295a). When the force increases and attains a certain critical value P_{cr} the bar buckles and takes the form depicted in Fig. 295b. Let us denote the transverse deviation of a point x of the bar from its original position by y. As it is proved in courses on strength of materials, the function $y(x)$ satisfies, within a sufficient accuracy, the differential equation and the boundary conditions
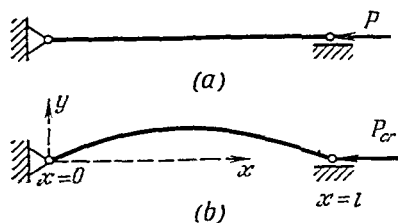
$$y'' + \frac{P}{EJ}y = 0, \quad y(0) = y(l) = 0 \tag{101}$$

where E is the so-called **modulus of elasticity** (also called **Young's modulus** after T. Young, 1773-1829, an English physicist, physician and astronomer, one of the creators of the wave theory of light) and J is the **moment of inertia** of the cross section of the bar. As follows from (100), there is a basic case for problem (101) when

$$\frac{P}{EJ} < \left(\frac{\pi}{l}\right)^2 \tag{102}$$

This means that if condition (102) is fulfilled the problem has only the zero solution, that is there is no buckling in this case. When inequality (102) turns into the equality, as P increases, there appears a singular case. Then, besides the zero solution, problem (101) possesses solutions of the form $y = C \sin \frac{\pi}{l} x$ where C is an arbitrary constant. In this case there are no forces that can keep the bar in the rectilinear equilibrium state and therefore even arbitrarily small external actions can result in finite deviations from this state. This means that the bar becomes unstable. The expression

$$P_{cr} = EJ \left(\frac{\pi}{l}\right)^2$$

of the critical force was found by Euler in 1757. One may think that after a further increase of the force, when we have $P > P_{cr}$, the bar will again become rectilinear but this conclusion is incorrect. The fact is that equation (101) is applicable only to the limiting case of small deviations from the rectilinear state. A more comprehensive investigation of the exact non-linear equation describing the state of the bar for any deviations shows that after P exceeds P_{cr} there appears a new curvilinear equilibrium state which is stable and which exists together with the unstable rectilinear equilibrium state. The curvature of the curvilinear equilibrium form rapidly increases, as P increases, and finally the bar breaks.

The notion of an influence function (Green's function) introduced in Sec. XIV.26 can be applied for solving the non-homogeneous equation with the homogeneous boundary conditions

$$y'' + p(x) y' + q(x) y = f(x) \quad (a \leqslant x \leqslant b),$$
$$y(a) = 0, \quad y(b) = 0 \tag{103}$$

in the basic (non-singular) case. Indeed, we can interpret the function $f(x)$ as an "external" action. Then $y(x)$ is interpreted as a result of the action, i.e. $y(x) = \tilde{f}(x)$. The superposition principle is also applicable here (why is it so?).

According to Sec. XIV.26, let us denote the solution of problem (103) in which the delta function $\delta(x - \xi)$ substitutes for $f(x)$ by $G(x, \xi)$; then the solution of problem (103) for an arbitrary function $f(x)$ is expressed in the form

$$y(x) = \int_a^b f(\xi) G(x, \xi) \, d\xi \tag{104}$$

We now take a simple example. Consider the problem

$$y'' = f(x) \quad (0 \leqslant x \leqslant l), \quad y(0) = y(l) = 0 \tag{105}$$

If we substitute $\delta(x - \xi)$ for $f(x)$ we simply obtain $y'' = 0$ for $0 \leqslant x < \xi$ and for $\xi < x \leqslant l$. Thus the solution is

$$y = ax + b \quad \text{for} \quad 0 \leqslant x < \xi \quad \text{and}$$
$$y = cx + d \quad \text{for} \quad \xi < x \leqslant l$$

where a, b, c and d are some constants. The boundary conditions imply that $b = 0$ and $cl + d = 0$, i.e.

$$y = ax \quad \text{for } 0 \leqslant x < \xi \quad \text{and} \tag{106}$$
$$y = c(x - l) \quad \text{for } \xi < x \leqslant l$$

If we integrate the equality $y'' = \delta(x - \xi)$ from $x = \xi - 0$ to $x = \xi + 0$ we get $y'(\xi + 0) - y'(\xi - 0) = 1$. By the way, the result would be the same for the left-hand side of equation (103) because integrating a finite function over an interval of zero length results in zero. The repeated integration of the delta function yields a continuous function (see Sec. XIV.25) and therefore we have $y(\xi - 0) = y(\xi + 0)$. Hence we obtain $c - a = 1$ and $a\xi = c(\xi - l)$ from (106). It follows that
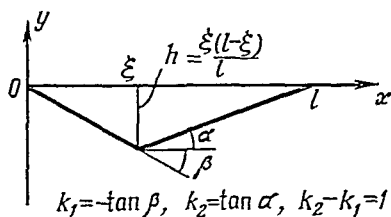
$$h = \frac{\xi(l - \xi)}{l}$$

$$k_1 = -\tan\beta, \quad k_2 = \tan\alpha, \quad k_2 - k_1 = 1$$

Fig. 296

$$a = -\frac{l - \xi}{l}, \qquad c = \frac{\xi}{l}$$

Substituting the values of a and c into (106) we find the influence function of problem (105):

$$G(x, \xi) = \begin{cases} -\dfrac{(l - \xi)x}{l} & \text{for } 0 \leqslant x < \xi \\[2mm] -\dfrac{\xi(l - x)}{l} & \text{for } \xi < x \leqslant l \end{cases}$$

The function is shown in Fig. 296. On the basis of formula (104) we obtain the formula of the solution of problem (105) valid for any function $f(x)$:

$$y = \int_0^l G(x, \xi) f(\xi)\,d\xi = \int_0^x G(x, \xi) f(\xi)\,d\xi +$$
$$+ \int_x^l G(x, \xi) f(\xi)\,d\xi = -\frac{(l - x)}{l} \int_0^x \xi f(\xi)\,d\xi -$$
$$- \frac{x}{l} \int_x^l (l - \xi) f(\xi)\,d\xi$$

§ 5. Linear Equations with Constant Coefficients

Linear equations with constant coefficients form an important class of differential equations whose integration can be completed in a comparatively simple way.

17. Homogeneous Equations. For definiteness, we now consider the third-order equation

$$z''' + a_1 z'' + a_2 z' + a_3 z = 0 \qquad (107)$$

where a_1, a_2 and a_3 are constants. Euler's idea is to seek a particular solution of the equation in the form

$$z = e^{px} \qquad (108)$$

where p is a constant that must be appropriately chosen.

Substituting (108) into (107) we see that

$$e^{px} (p^3 + a_1 p^2 + a_2 p + a_3) = 0$$

The first factor being unequal to zero, we obtain

$$p^3 + a_1 p^2 + a_2 p + a_3 = 0 \qquad (109)$$

Thus, a function of form (108) satisfies equation (107) if and only if p satisfies equation (109). Algebraic equation (109) in the unknown quantity p is called the **characteristic equation** of equation (107). The left-hand side of the characteristic equation (109) is called the **characteristic polynomial** of equation (107). The degree of a characteristic equation is equal to the order of the corresponding equation.

Equation (109) has three roots (see Sec. VIII.8) which we denote as p_1, p_2 and p_3. There can be different cases here, namely, the following ones.

1. Let all the roots be real and simple (that is distinct from each other). Then, by formula (108), we have three particular solutions of equation (107) of the form

$$z_1 = e^{p_1 x}, \quad z_2 = e^{p_2 x} \quad \text{and} \quad z_3 = e^{p_3 x}$$

The solutions being linearly independent (i.e. none of them being a linear combination of the rest), the general solution of equation (107), as it is implied by Sec. 14. has the form

$$z = C_1 e^{p_1 x} + C_2 e^{p_2 x} + C_3 e^{p_3 x} \qquad (110)$$

2. Let all the roots be simple but let there be imaginary roots among them. Then there appear complex functions of a real argument on the right-hand side of formula (110) (see Sec. VIII.6). But it is easy to show that the whole theory of linear equations (see § 4) can be automatically extended to the case when all the coefficients are complex functions or numbers and all the solutions are complex

functions of a real argument. Formula (110) can therefore be applied to the case of imaginary simple roots of equation (109). It is apparent that the arbitrary constants are also complex here in the general case.

But when we consider real functions it is also preferable to obtain the answer in the real form. To attain this we can take advantage of the following assertion: if a homogeneous linear equation with real coefficients has a complex particular solution the real part and the imaginary part of the solution are solutions of the same equation. Actually, if $L[y_1 + iy_2] = 0$ (see Sec. 14 on the notation) then $L[y_1] + iL[y_2] = 0$ from which we deduce $L[y_1] = 0$ and $L[y_2] = 0$ (why?).

Hence, if the coefficients of equation (107) are real and if it has a particular solution

$$e^{(r+is)x} = e^{rx} \cos sx + ie^{rx} \sin sx$$

[see formula (VIII.12)] then the functions

$$e^{rx} \cos sx \quad \text{and} \quad e^{rx} \sin sx \tag{111}$$

are also solutions of equation (107). Since complex roots of an algebraic equation with real coefficients form conjugate pairs of complex numbers (see Sec. VIII.8), in our case 2 there are two roots of the form

$$p_1 = r + is \quad \text{and} \quad p_2 = r - is$$

whereas the third root p_3 is real. Hence, the general solution can be written in the form

$$z = C_1 e^{rx} \cos sx + C_2 e^{rx} \sin sx + C_3 e^{p_3 x} \tag{112}$$

instead of (110).

For instance, taking equation (88) which describes free oscillations we obtain the characteristic equation $p^2 + \omega_0^2 = 0$ with the roots $p_{1,2} = \pm i\omega_0 = 0 \pm i\omega_0$. Thus, the general solution of the equation is analogous to formula (112):

$$z = C_1 e^{0 \cdot t} \cos \omega_0 t + C_2 e^{0 \cdot t} \sin \omega_0 t$$

Thus, we have obtained the solution of the equation in form (89).

Formula (112) is sometimes rewritten in another form by taking advantage of the formula

$$C_1 \cos sx + C_2 \sin sx = M \sin (sx + \alpha)$$

To obtain this expression we must put

$$C_1 = M \sin \alpha, \quad C_2 = M \cos \alpha, \quad M = \sqrt{C_1^2 + C_2^2} \quad \text{and} \quad \tan \alpha = \frac{C_1}{C_2}$$

(compare with Sec. I.29). Then, instead of (112), we get

$$z = M e^{rx} \sin (sx + \alpha) + C_3 e^{p_3 x} \tag{113}$$

where the role of arbitrary constants is played by M, α and C_3.

3. Let there be multiple roots among the roots of characteristic equation (109). For example, let $p_2 = p_1$ and $p_3 \neq p_1$. Then, of course, formula (110) does not yield the general solution since formula (108) gives only two distinct solutions in this case (i.e. $e^{p_1 x}$ and $e^{p_3 x}$).

To find a third solution we begin with the case when $p_2 = p_1 + \Delta p$ where $|\Delta p| \neq 0$ is small. Then equation (107) has the solution

$$e^{p_2 x} = e^{p_1 x} e^{\Delta p \cdot x} = e^{p_1 x} \left(1 + \Delta p \cdot x + \frac{(\Delta p)^2 x^2}{2!} + \cdots \right)$$

together with the solution $e^{p_1 x}$ [here we have used expansion (IV.55)]. Therefore the linear combinations

$$e^{p_2 x} - e^{p_1 x} = e^{p_1 x} \left(\Delta p \cdot x + \frac{(\Delta p)^2 x^2}{2!} + \cdots \right)$$

and

$$\frac{e^{p_2 x} - e^{p_1 x}}{\Delta p} = e^{p_1 x} \left(x + \frac{\Delta p \cdot x^2}{2!} + \cdots \right) \tag{114}$$

are also solutions of the equation.

After dividing by Δp we can pass to the limit, as $\Delta p \to 0$. Then all the terms containing Δp tend to zero, and therefore we obtain the solution $x e^{p_1 x}$ for $\Delta p = 0$. that is for $p_2 = p_1$. Hence, in this case equation (107) has the general solution of the form

$$z = C_1 e^{p_1 x} + C_2 x e^{p_1 x} + C_3 e^{p_3 x}$$

Similarly, in the case $p_1 = p_2 = p_3$ equation (107) has the solution $e^{p_1 x}$ and, besides, the solutions $x e^{p_1 x}$ and $x^2 e^{p_1 x}$. To prove this we can take the second divided difference instead of (114) and pass to the limit (see Sec. V.7). Therefore the general solution for this case has the form

$$z = C_1 e^{p_1 x} + C_2 x e^{p_1 x} + C_3 x^2 e^{p_1 x}$$

Equations of an arbitrary order are investigated in a similar way. If a root p of a characteristic equation has a multiplicity k then the functions

$$e^{px}, \quad x e^{px}, \quad \ldots, \quad x^{k-1} e^{px}$$

are particular solutions of the corresponding differential equation. If a pair of complex conjugate roots $r \pm is$ is of multiplicity k then the functions

$$e^{rx} \cos sx, \quad e^{rx} \sin sx, \quad x e^{rx} \cos sx, \quad x e^{rx} \sin sx, \ldots$$
$$\ldots, \quad x^{k-1} e^{rx} \cos sx, \quad x^{k-1} e^{rx} \sin sx$$

are particular solutions.

Thus, the only practical difficulty in integrating a homogeneous linear equation with constant coefficients lies in solving the corres-

ponding characteristic equation which can be performed by means of the methods described in Secs. V.2-5 and VIII.9.

As an example, we now consider free oscillations of a material point when there is a linear law of elasticity and an additional force of viscous friction which is proportional to the first degree of the velocity. In this case we must add the term $-f\frac{dy}{dt}$ to the right-hand side of equation (3) where f is the coefficient of viscous friction.
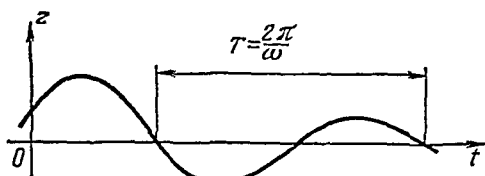
Fig. 297

After transposing all the terms to the left-hand side and dividing by M we obtain the equation

$$z'' + 2hz' + \omega_0^2 z = 0 \qquad (115)$$

similar to (88) where $2h = \frac{f}{M}$ and $\omega_0^2 = \frac{k}{M}$.

In solving the characteristic equation

$$p^2 + 2hp + \omega_0^2 = 0 \qquad (116)$$

we can encounter two basic cases. If $h < \omega_0$, that is if the friction is comparatively small, equation (116) has the solution

$$p_{1,2} = -h \pm \sqrt{h^2 - \omega_0^2} = -h \pm i\sqrt{\omega_0^2 - h^2}$$

and the general solution of equation (115) is therefore analogous to (113):

$$z(t) = Me^{-ht}\sin(\omega t + \alpha)$$

where $\omega = \sqrt{\omega_0^2 - h^2}$.

We see that the presence of small friction yields damped oscillations whose amplitude decreases according to an exponential law (because of the factor e^{-ht}). Besides, the friction decreases the frequency (since $\omega < \omega_0$). The graph of the solution is depicted in Fig. 297. Zeros of the solution are defined by the factor $\sin(\omega t + \alpha)$ and therefore they are equally spaced. After the time period $T = \frac{2\pi}{\omega}$ elapses the sine repeats its value and the function e^{-ht} "educes" the factor $e^{-h\frac{2\pi}{\omega}}$ which causes damping. The time interval T is often

called the "period" of the oscillations although the function $z(t)$ is a non-periodic function here because during each subsequent "period" the amplitude of z decreases $e^{-h\frac{2\pi}{\omega}}$ times (whereas the general form of the function does not change).

If $h > \omega_0$, that is if the friction is comparatively large, equation (116) has real roots, and equation (115) has the general solution

$$z(t) = C_1 e^{p_1 t} + C_2 e^{p_2 t} = C_1 e^{-(h-\sqrt{h^2-\omega_0^2})\,t} + C_2 e^{-(h+\sqrt{h^2-\omega_0^2})\,t}$$

Only the first summand is essential here for large t (why?). Hence, we obtain an exponential damping without oscillations in this case. This is the so-called *aperiodic damping*.

Theoretically, there is a possibility of a "bordering" case when $h = \omega_0$. Then equation (116) has a double root. We leave it to the reader to verify that in this case we also have aperiodic damping.

18. **Non-Homogeneous Equations with Right-Hand Sides of Special Form.** We now turn to non-homogeneous linear equations with constant coefficients. For instance, let us take a third-order equation of the form

$$y''' + a_1 y'' + a_2 y' + a_3 y = f(x) \quad (a_1,\ a_2,\ a_3 = \text{const}) \tag{117}$$

The corresponding homogeneous equation being always solvable (see Sec. 17), the considerations of Sec. 15 imply that it is only a single particular solution of equation (117) that we have to find. For the right-hand side of the general form this can be achieved by the method of variation of arbitrary constants (see Sec. 15). But there exists an important and a rather wide class of right-hand sides of a special form for which a particular solution can be found considerably faster and simpler by applying the method of undetermined coefficients.

We first take the equation

$$y''' + a_1 y'' + a_2 y' + a_3 y = K e^{\lambda x} \quad (K,\ \lambda = \text{const}) \tag{118}$$

It is natural to seek a particular solution of the equation in the form

$$y = A e^{\lambda x} \tag{119}$$

where the constant A is yet unknown. Substituting (119) into (118) we obtain

$$A\lambda^3 e^{\lambda x} + A a_1 \lambda^2 e^{\lambda x} + A a_2 \lambda e^{\lambda x} + A a_3 e^{\lambda x} = K e^{\lambda x}$$

This results in

$$A = \frac{K}{\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3} \quad \text{and} \quad y = \frac{K}{P(\lambda)} e^{\lambda x} \tag{120}$$

(after cancelling out the factor $e^{\lambda x}$) where $P(\lambda)$ designates the characteristic polynomial.

The result is valid if $P(\lambda) \neq 0$, that is if λ is not a root of the characteristic equation. If $P(\lambda) = 0$ then function (119) satisfies homogeneous equation (107) and therefore it cannot satisfy equation (118).

Let $P(\lambda) = 0$ and $P'(\lambda) \neq 0$ which means that λ is a simple root (see Sec. VIII.8) of the characteristic equation. Then, following the same way of reasoning as in Sec. 17 when we considered the case of multiple roots, we substitute $\lambda_1 = \lambda + \alpha$ for λ into the right-hand side of equation (118) where $|\alpha|$ is small but different from zero. The value λ_1 is no longer a root of the characteristic equation and therefore formula (120) suggests that equation (118) has a particular solution of the form

$$\frac{K}{P(\lambda_1)} e^{\lambda_1 x} = \frac{K}{P(\lambda + \alpha)} e^{(\lambda + \alpha) x} = \frac{K}{P(\lambda + \alpha)} e^{\lambda x} \left(1 + \alpha x + \frac{\alpha^2 x^2}{2!} + \dots \right) =$$
$$= \frac{K}{P(\lambda + \alpha)} e^{\lambda x} + K \frac{\alpha}{P(\lambda + \alpha)} e^{\lambda x} \left(x + \frac{\alpha x^2}{2!} + \dots \right)$$

The first summand on the right-hand side of the last relation satisfying the corresponding homogeneous equation, the second summand is also a particular solution of equation (118) in which $\lambda = \lambda_1$ (why is it so?). We now pass to the limit in the second summand as $\alpha \to 0$. Applying L'Hospital's rule to calculating the limit $\lim\limits_{\alpha \to 0} \frac{\alpha}{P(\lambda + \alpha)}$ we obtain the expression

$$y = \frac{K}{P'(\lambda)} x e^{\lambda x} \tag{121}$$

(check it up!) which is a particular solution of equation (118) for the original value of λ.

In a similar way we investigate the case when λ is a double root of the characteristic equation and find that in this case there is a particular solution of equation (118) of the form

$$y = \frac{K}{P''(\lambda)} x^2 e^{\lambda x}$$

and so on.

By means of a similar but a little more complicated procedure we can prove that the equation

$$y''' + a_1 y'' + a_2 y' + a_3 y = Q_m(x) e^{\lambda x} \tag{122}$$

where $Q_m(x)$ is a given polynomial of degree m possesses a particular solution of the form

$$y = R_m(x) e^{\lambda x} \tag{123}$$

provided λ is not a root of the characteristic equation [$R_m(x)$ denotes some other polynomial of degree m]. The solution can also be found by means of the method of undetermined coefficients. To apply the method we write expression (123) with literal coefficients and substitute it into (122). Then the usual procedure of equating coeffi-

cients in similar terms, based on the identical equality of both sides of the relation, yields the numerical values of the coefficients entering into (123). Similarly, if λ is a root of the characteristic equation of multiplicity k then there is a particular solution of the form

$$y = x^k R_m (x) e^{\lambda x}$$

The case $\lambda = 0$ is not excluded here and therefore we can also obtain a particular solution when there is a polynomial of degree m on the right-hand side of equation (122) (without an exponential factor).

We can also consider the equation

$$y''' + a_1 y'' + a_2 y' + a_3 y = Q_m (x) e^{\mu x} \cos \nu x$$

Since we can rewrite the right-hand side in the form

$$Q_m (x) e^{\mu x} \frac{e^{i\nu x} + e^{-i\nu x}}{2} = \frac{Q_m (x)}{2} e^{(\mu + i\nu) x} + \frac{Q_m (x)}{2} e^{(\mu - i\nu) x}$$

(see Sec. VIII.4), formula (123) suggests that if $\lambda = \mu \pm i\nu$ is not a root of the characteristic equation, a particular solution can be sought for in the form

$$y = R_m (x) e^{(\mu + i\nu)x} + \widetilde{R}_m (x) e^{(\mu - i\nu)x} =$$
$$= R_m (x) e^{\mu x} (\cos \nu x + i \sin \nu x) + \widetilde{R}_m (x) e^{\mu x} (\cos \nu x - i \sin \nu x) =$$
$$= [R_m (x) + \widetilde{R}_m (x)] e^{\mu x} \cos \nu x + [i R_m (x) - i \widetilde{R}_m (x)] e^{\mu x} \sin \nu x$$

Changing the notation we arrive at a particular solution of the form

$$y = T_m (x) e^{\mu x} \cos \nu x + S_m (x) e^{\mu x} \sin \nu x \qquad (124)$$

where $T_m (x)$ and $S_m (x)$ are polynomials of degree m which can be found with the help of the method of undetermined coefficients. In the case the right-hand side is of the form

$$Q_m (x) e^{\mu x} \sin \nu x \quad \text{or} \quad Q_m (x) e^{\mu x} \cos (\nu x + \alpha) \quad \text{or}$$
$$Q_m (x) e^{\mu x} \sin (\nu x + \alpha) \quad (\alpha = \text{const})$$

we can also seek a particular solution in form (124).

If $\lambda = \mu \pm i\nu$ is a root of the characteristic equation of multiplicity k the right-hand side of formula (124) should be additionally multiplied by x^k.

For instance, let us consider equation (87) with a sinusoidal right-hand member $K \sin \omega t$:

$$y'' + \omega_0^2 y = K \sin \omega t \qquad (125)$$

The equation describes forced oscillations under the action of an external sinusoidal force of frequency ω.

According to formula (124). if $\lambda = \pm i\omega$ does not coincide with any root of the characteristic equation, i.e. if $\omega \neq \omega_0$, a particular

solution can be found in the form

$$y = A \cos \omega t + B \sin \omega t$$

Substituting this expression into equation (125) we obtain

$$-A\omega^2 \cos \omega t - B\omega^2 \sin \omega t + A\omega_0^2 \cos \omega t + \\ + B\omega_0^2 \sin \omega t = K \sin \omega t$$

The last equality being an identity, we must have

$$-A\omega^2 + A\omega_0^2 = 0 \quad \text{and} \quad -B\omega^2 + B\omega_0^2 = K$$

Hence, $A = 0$ and $B = \dfrac{K}{\omega_0^2 - \omega^2}$, and thus we obtain

$$y = \frac{K}{\omega_0^2 - \omega^2} \sin \omega t \qquad (126)$$

To obtain the general solution of equation (125) we must add the general solution of equation (88) to expression (126). Consequently, if the frequency of the external force is different from the fundamental frequency of oscillations we have a superposition of two harmonic oscillations of forms (89) and (126). Function (126) describes the so-called *forced oscillations* whose frequency coincides with the frequency of the external action and whose amplitude and phase have completely specified values. Function (89), which is a solution of equation (88), describes free oscillations whose amplitude and phase depend on the initial data.

Formula (126) shows that when $\omega \neq \omega_0$ and is close to ω_0 the amplitude of forced oscillations becomes very large. But if $\omega = \omega_0$ then, according to the general theory, a particular solution of equation (125) can be found in the form $y = At \cos \omega_0 t + Bt \sin \omega_0 t$.

The substitution of the last expression into the equation results in the formula

$$y = -\frac{K}{2\omega_0} t \cos \omega_0 t$$

which can also be deduced from (126) by analogy with formula (121). (Let the reader verify the validity of the above expression.)

We see that if the frequency of a harmonic external action is equal to the fundamental frequency of oscillations the amplitude of forced oscillations increases according to a linear law. This important phenomenon is well known in physics and engineering and is called resonance.

19. Euler's Equations. Euler's equations are of the form

$$(ax + b)^n y^{(n)} + a_1 (ax + b)^{n-1} y^{(n-1)} + \ldots + \\ + a_{n-1} (ax + b) y' + a_n y = f(x)$$

where $a_1, a_2, \ldots, a_n$ are constant coefficients.

Euler's equation can be easily reduced to a linear equation with constant coefficients by means of the change of the independent variable

$$| ax + b | = e^t, \quad \text{i.e.} \quad t = \ln | ax + b |$$

For the sake of simplicity, let us suppose that $ax + b > 0$ and take a homogeneous second-order equation:

$$(ax + b)^2 y'' + a_1 (ax + b) y' + a_2 y = 0 \qquad (127)$$

After the independent variable is changed we obtain

$$ax + b = e^t, \qquad t = \ln (ax + b), \qquad (128)$$

$$y' = \frac{dy}{dx} = \frac{dy}{dt} \cdot \frac{dt}{dx} = \frac{dy}{dt} \cdot ae^{-t} \quad \text{and}$$

$$y'' = \frac{dy'}{dx} = \frac{dy'}{dt} \cdot \frac{dt}{dx} = \left(\frac{d^2 y}{dt^2} ae^{-t} - \frac{dy}{dt} ae^{-t} \right) ae^{-t}$$

Substituting the results into equation (127) we receive

$$a^2 \left(\frac{d^2 y}{dt^2} - \frac{dy}{dt} \right) + a_1 a \frac{dy}{dt} + a_2 y = 0 \qquad \cdot$$

This is an equation with constant coefficients which can be solved by means of the methods described in Sec. 17, i.e. we must put

$$y = e^{pt} \qquad (129)$$

Then we solve the corresponding characteristic equation etc. and, finally, turn back from t to x.

It is possible to avoid substitution (128) because (128) and (129) imply that

$$y = (ax + b)^p \qquad (130)$$

We can therefore directly substitute (130) into (127), i.e. we can seek a solution in form (130). This yields a characteristic equation for determining p, the degree of the equation being equal to the order of equation (127). But one must take into account that, in case there are multiple roots of the characteristic equation, equation (127) possesses solutions of the form

$$y = te^{pt} = (ax + b)^p \ln (ax + b)$$

besides solutions of form (130) (the form of these additional solutions depends on the multiplicity of the root; see case 3 in Sec. 17).

20. Operators and the Operator Method of Solving Differential Equations. The notion of an operator was introduced in Sec. XIV.26. We also considered some examples of operators including the operator of differentiation D. In Sec. 14 we introduced a differential

operator L. Further examples of operators are the **shift operator** T and the **difference operator** Δ which are defined by the formulas

$$Tf(x) = f(x + h) \quad \text{and} \quad \Delta f(x) = f(x + h) - f(x) \qquad (131)$$

where h is a given step. We also introduce the **operator of multiplication by a number** C which we denote by the same letter C (including the **unit operator** 1 which does not change a function and the **zero operator** 0 which transforms any function into the function which is identically equal to zero). There is also an **operator of multiplication by a given function** and so on.

Operators can be added together and multiplied by numbers according to the following rule which looks quite natural: if A and B are operators and α is a number then

$$(A + B) f \equiv Af + Bf \quad \text{and} \quad (\alpha A) f \equiv \alpha (Af)$$

For instance, equality (131) suggests that

$$\Delta = T - 1 \quad \text{and} \quad T = 1 + \Delta$$

All the axioms of linear operations hold for these rules of addition of operators and multiplication by a number (see Sec. VII.17).

We can multiply an operator by another one according to Sec. XI.6: if A and B are operators then AB is a new operator defined by the formula

$$(AB) f \equiv A (Bf)$$

which means that to obtain ABf we must first apply the operator B to the function f and then apply operator A to the result. We can easily verify the following rules:

$$A (BC) = (AB) C \quad \text{and} \quad (\alpha A + \beta B) C = \alpha AC + \beta BC \qquad (132)$$

where $\alpha. \beta = \text{const}$. But the equality $AB = BA$ may not hold, that is in the general case the result of performing two operations may depend on the order they are performed. But nevertheless in particular cases we can have $AB = BA$ and then we say that the operators *commute* with each other. For example, all the above operators D, T, Δ and C commute with each other because we have

$$DTf(x) = D(Tf(x)) = D(f(x + h)) = f'(x + h),$$
$$TDf(x) = T(f'(x)) = f'(x + h) \quad \text{etc.}$$

On the other hand, the operator of differentiation of a function and the operator of multiplication of a function by a given function do not commute (the reader should verify it!).

The first property (132) enables us to write ABC instead of (AB) C or A (BC). Thus, ABC is an operator whose action upon an object reduces to performing the operations C, B and A in succession. An

operator of the form ABCD is defined similarly and so on. If we take equal factors then we arrive at *powers of an operator*: A^2, A^3, A^4 etc. Hence, A^n designates the repeated application of the operator A. For instance, we have $D^2 f = f''$, and $\Delta^2 f$ is the second difference (see Sec. XI.6) etc.

An operator A is called a **linear operator** if

$$A (f_1 + f_2) = A (f_1) + A (f_2) \quad \text{and} \quad A (\alpha f) = \alpha A (f) \qquad (133)$$

where $\alpha = \text{const}$ (see Sec. XI.6). It is natural to interpret the first property as the principle of superposition, and the second property can be deduced from the first (see Sec. XIV.26). Even in the case when the explicit expression of an operator is unknown the validity of the superposition principle indicates the linearity of the operator which enables us to draw some useful conclusions; for example, we can speak about its influence function (see Sec. XIV.26) in such a case.

Both properties (133) can be put down as

$$A (\alpha f_1 + \beta f_2) = \alpha A f_1 + \beta A f_2$$

where α and β are constants. It is easy to verify the following property of a linear operator A:

$$A (\alpha B + \beta C) = \alpha A B + \beta A C \qquad (134)$$

where B and C are arbitrary operators. To deduce the property one should apply both parts of (134) to an arbitrary function f and show that the results will be the same, namely, equal to $\alpha A (Bf) + \beta A (Cf)$.

All the operators we have taken here as examples are linear. Indeed, as we know, the derivative of a sum equals the sum of the derivatives and so on. An example of a non-linear operator is the operator of squaring a function or the operator of forming the absolute value of a function (verify that the operators are non-linear!) and the like. In performing linear operations on linear operators and the operation of multiplying operators by each other we can use rules of elementary algebra but at the same time we must pay attention to the order of factors. For instance, $(A + B)^2 = (A + B) (A + B) = A^2 + AB + BA + B^2$. This assertion is suggested by (132) and (134).

In addition, if the operators commute with each other the order of factors does not matter either. For example, in such a case we have $(A + B)^2 = A^2 + 2AB + B^2$ and so on.

We can also consider power series (see Sec. IV.16) of operators. For instance, we can define the operator e^A as

$$e^A = 1 + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots \qquad (135)$$

and so on.

Generally, such a series or any other operator cannot be applied to all functions because usually there is a certain class of functions for which an operator makes sense, and thus it should be used only for these functions.

As for example (135), we see that the greater the number of the terms taken, the more accurate the result. Theoretically, the exact result is obtained only in the limiting process. From the practical point of view this means that the number of terms must be sufficiently large for the result to be regarded as exact.

Taylor's series

$$f(x+h) = f(x) + \frac{f'(x)}{1!}h + \frac{f''(x)}{2!}h^2 + \ldots$$

see Sec. IV.16) can be written in the form

$$Tf = \left(1 + \frac{D}{1!}h + \frac{D^2}{2!}h^2 + \ldots\right)f = e^{hD}f$$

This implies a relation between the operators T, Δ and D, namely

$$T = e^{hD} \quad \text{and} \quad \Delta = e^{hD} - 1$$

The inverse formula

$$D = \frac{1}{h}\ln T = \frac{1}{h}\ln(1+\Delta) = \frac{1}{h}\left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \ldots\right)$$

[which corresponds to formula (IV.61) of expansion of the natural logarithm] is nothing but formula (V.32) of numerical differentiation.

We can consider an **operator equation** of the form

$$Ay = f \tag{136}$$

where f is a given function and the function y is sought for. If there is a solution of equation (136) it is natural to denote it as $y = A^{-1}f$. In case the operator is linear the equation is also said to be a linear equation. We can immediately extend properties 1-3 in Sec. 14 and property 1 in Sec. 15 to general linear equations. But it should be taken into account that there are cases when a homogeneous equation has infinitely many linearly independent solutions and cases when a non-homogeneous equation has no solutions.

We shall demonstrate the application of the operator of differentiation to solving linear equations with constant coefficients of form (117) (the so-called *operator method* of solving equations). The equation can be rewritten as

$$(D^3 + a_1D^2 + a_2D + a_3)y = f(x)$$

There is a linear differential operator of the third order with constant coefficients inside the parentheses. We can factor the operator into linear factors according to algebraic rules (see Sec. VIII.8):

$$(D - p_1)(D - p_2)(D - p_3)y = f(x) \tag{137}$$

where p_1, p_2 and p_3 are the roots of the characteristic equation (see Sec. 17). Since

$$D\,(e^{-px}y) = (e^{-px}y)' = -pe^{-px}y + e^{-px}y' =$$
$$= e^{-px}\,(y' - py) = e^{-px}\,(D - p)\,y$$

we have

$$(D - p)\,y = e^{px}D\,(e^{-px}y)$$

and therefore equation (137) can be rewritten in the form

$$e^{p_1x}D\,(e^{-p_1x}\,e^{p_2x}\,D\,(e^{-p_2x}\,e^{p_3x}\,D\,(e^{-p_3x}\,y))) = f\,(x)$$

Transposing the factors from the left-hand side to the right-hand side and taking into account that the equality $Dy = z$ is equivalent to the equality $y = \int z\,dx$ we obtain the general solution of the original equation:

$$y = e^{p_3x}\int e^{(p_2-p_3)\,x}\left(\int e^{(p_1-p_2)\,x}\left(\int e^{-p_1x}f\,(x)\,dx\right)dx\right)dx$$

It is apparent that after integration we obtain the same result as in Secs. 15 and 18 although we have used a different approach. There are more complicated problems for which the operator method may be essentially useful. We must note that such a simple factorization of an operator into a product of several operators of lower order can rarely be applied effectively to linear differential operators with variable coefficients and moreover to non-linear operators.

We now consider one more simple example. Let it be necessary to find a solution of the equation

$$y'' + \omega^2 y = f\,(x)$$

where all the quantities are regarded as being real. We write, in succession,

$$(D^2 + \omega^2)\,y = f\,(x), \quad (D - i\omega)\,(D + i\omega)\,y = f\,(x),$$
$$e^{i\omega x}\,D\,(e^{-i\omega x}\,(D + i\omega)\,y) = f\,(x),$$

$$(D + i\omega)\,y = e^{i\omega x}\int e^{-i\omega x}f\,(x)\,dx \quad\text{and}\quad \omega y = \operatorname{Im}\left(e^{i\omega x}\int e^{-i\omega x}f\,(x)\,dx\right)$$

(the sign Im designates the imaginary part according to the notation introduced in Secs. VIII.1 and VIII.6). Finally,

$$y = \frac{1}{\omega}\operatorname{Im}\left(e^{i\omega x}\int e^{-i\omega x}f\,(x)\,dx\right)$$

§ 6. Systems of Linear Equations

21. **Systems of Linear Equations.** For definiteness, let us consider a homogeneous linear system of three first-order equations in three unknown functions $y\,(x)$, $z\,(x)$ and $u\,(x)$. We take the system in the

form solved for the derivatives of the functions:

$$\left.\begin{array}{l} y' = a_1(x)\, y + b_1(x)\, z + c_1(x)\, u \\ z' = a_2(x)\, y + b_2(x)\, z + c_2(x)\, u \\ u' = a_3(x)\, y + b_3(x)\, z + c_3(x)\, u \end{array}\right\} \quad (138)$$

We remind the reader that a system of any order can be reduced to a first-order system (see Sec. 11) and that the operation of resolving a system with respect to the derivatives is performed algebraically without solving the differential equations.

We can easily pass from system (138) to an equivalent third-order equation (see Sec. 11) which is also a homogeneous linear equation. Therefore all the properties of homogeneous linear equations enumerated in Sec. 14 are extended to system (138). The sum of two solutions $y = y_1$, $z = z_1$, $u = u_1$ and $y = y_2$, $z = z_2$, $u = u_2$ should be understood as a new solution of the form $y = y_1 + y_2$, $z = z_1 + z_2$, $u = u_1 + u_2$. Similarly, the product of a solution $y = y_1$, $z = z_1$, $u = u_1$ by a constant C is the new solution $y = Cy_1$, $z = Cz_1$, $u = Cu_1$. Hence, linear operations on solutions are performed here in the same way as on vectors (see Sec. VII.10).

In particular, the general solution of system (138) has the form

$$\left.\begin{array}{l} y = C_1 y_1 + C_2 y_2 + C_3 y_3 \\ z = C_1 z_1 + C_2 z_2 + C_3 z_3 \\ u = C_1 u_1 + C_2 u_2 + C_3 u_3 \end{array}\right\} \quad (139)$$

where C_1, C_2 and C_3 are arbitrary constants and (y_1, z_1, u_1), (y_2, z_2, u_2) and (y_3, z_3, u_3) are three linearly independent solutions of system (138), that is such solutions that none of them is a linear combination of the rest.

Let us dwell in more detail on property 4 in Sec. 14. If a nonzero solution (y_1, z_1, u_1) of system (138) is known then making the substitutions $y = y_1 \bar{y}$, $z = z_1 \bar{z}$, $u = u_1 \bar{u}$ and $\bar{y} = \bar{u} + v$, $\bar{z} = \bar{u} + w$ we easily derive a system of two equations of the first order in two unknown functions $v(x)$ and $w(x)$ from which $\bar{u}$ is found by a single integration.

The verification of the last assertion is left to the reader.

All the properties enumerated in Sec. 15 are also extended to non-homogeneous linear systems of the form

$$\left.\begin{array}{l} y' = a_1(x)\, y + b_1(x)\, z + c_1(x)\, u + f_1(x) \\ z' = a_2(x)\, y + b_2(x)\, z + c_2(x)\, u + f_2(x) \\ u' = a_3(x)\, y + b_3(x)\, z + c_3(x)\, u + f_3(x) \end{array}\right\} \quad (140)$$

In particular, the method of variation of arbitrary constants (Sec. 3) is applied in the following manner. Let general solution (139) of the

corresponding homogeneous system (138) be known. Then a solution of system (140) is sought in the form

$$
\left.
\begin{array}{l}
y = \varphi_1(x) y_1 + \varphi_2(x) y_2 + \varphi_3(x) y_3 \\
z = \varphi_1(x) z_1 + \varphi_2(x) z_2 + \varphi_3(x) z_3 \\
u = \varphi_1(x) u_1 + \varphi_2(x) u_2 + \varphi_3(x) u_3
\end{array}
\right\}
$$

After substituting these expressions into (140) we obtain the system

$$
\left.
\begin{array}{l}
\varphi_1' y_1 + \varphi_2' y_2 + \varphi_3' y_3 = f_1(x) \\
\varphi_1' z_1 + \varphi_2' z_2 + \varphi_3' z_3 = f_2(x) \\
\varphi_1' u_1 + \varphi_2' u_2 + \varphi_3' u_3 = f_3(x)
\end{array}
\right\}
$$

(verify the calculations!). The derivatives φ_1', φ_2' and φ_3' are found algebraically from the last relations. Then, integrating, we obtain φ_1, φ_2 and φ_3.

Homogeneous linear systems with constant coefficients are especially important. For example, let us take the system

$$
\left.
\begin{array}{l}
y' = a_1 y + b_1 z + c_1 u \\
z' = a_2 y + b_2 z + c_2 u \\
u' = a_3 y + b_3 z + c_3 u
\end{array}
\right\}
\tag{141}
$$

in which all the coefficients a_1, b_1, . . ., c_3 are constant. The method of solving such a system is similar to that of Sec. 17. Namely, non-zero particular solutions are sought in the form

$$
y = \lambda e^{px}, \quad z = \mu e^{px}, \quad u = \nu e^{px}
\tag{142}
$$

where λ, μ, ν and p are unknown constants that must be determined. Substituting (142) into (141), cancelling out e^{px} and transposing all the terms to the left-hand side we obtain

$$
\left.
\begin{array}{l}
(a_1 - p) \lambda + b_1 \mu + c_1 \nu = 0 \\
a_2 \lambda + (b_2 - p) \mu + c_2 \nu = 0 \\
a_3 \lambda \quad b_3 \mu + (c_3 - p) \nu = 0
\end{array}
\right\}
\tag{143}
$$

These equalities should be regarded as a system of three homogeneous algebraic equations of the first degree in three unknowns λ, μ and ν. For the system to have a nonzero solution (it is the only solution that we are interested in, according to the end of Sec. VI.6), it is necessary and sufficient that the determinant of the system be equal to zero. Thus,

$$
\begin{vmatrix}
a_1 - p & b_1 & c_1 \\
a_2 & b_2 - p & c_2 \\
a_3 & b_3 & c_3 - p
\end{vmatrix} = 0
\tag{144}
$$

This is the characteristic equation of system (141) from which we find all the possible values of p.

Equation (144) being an algebraic equation of the third degree in p (why is it so?), there are three roots p_1, p_2 and p_3. If all the roots are simple we can substitute any of them into system (143) and find a nonzero solution λ, μ and ν. Then formula (142) yields the corresponding solution $y(x)$, $z(x)$, $u(x)$. The three particular solutions which correspond to $p = p_1$, $p = p_2$ and $p = p_3$ enable us to put down the general solution of system (141) in accord with formula (139):

$$\left.\begin{array}{l} y = C_1\lambda_1 e^{p_1 x} + C_2\lambda_2 e^{p_2 x} + C_3\lambda_3 e^{p_3 x} \\ z = C_1\mu_1 e^{p_1 x} + C_2\mu_2 e^{p_2 x} + C_3\mu_3 e^{p_3 x} \\ u = C_1\nu_1 e^{p_1 x} + C_2\nu_2 e^{p_2 x} + C_3\nu_3 e^{p_3 x} \end{array}\right\} \tag{145}$$

where C_1, C_2 and C_3 are arbitrary constants.

If equation (144) has imaginary roots we can retain the solution in form (145) (which will contain complex functions of x) or, by analogy with case 2 in Sec. 17, write the solution in the real form.

The case when equation (144) has multiple roots is more complicated. We shall not consider this case in the general form, but in concrete problems one can use the following procedure. For example, if p_1 is a double root particular solutions corresponding to $p = p_1$ should be looked for in the form

$$y = (\lambda x + \bar{\lambda})\, e^{p_1 x}, \quad z = (\mu x + \bar{\mu})\, e^{p_1 x},$$
$$u = (\nu x + \bar{\nu})\, e^{p_1 x} \tag{146}$$

in place of (142). Substituting (146) into (141), cancelling out $e^{p_1 x}$ and equating coefficients in the same powers of x we obtain equations from which we find two different and independent sets of possible values of the coefficients λ, $\bar{\lambda}$. . . ., $\bar{\nu}$. This results in two independent solutions of system (146). If a root of the characteristic equation is of higher multiplicity the corresponding particular solution of system (141) becomes more complicated.

Matrices are widely used in the theory of linear differential equations to simplify the form of writing such systems (see Secs. XI.1-2). To do this we usually rewrite system (138) in the form

$$\left.\begin{array}{l} y_1' = a_{11}(x)\, y_1 + a_{12}(x)\, y_2 + a_{13}(x)\, y_3 \\ y_2' = a_{21}(x)\, y_1 + a_{22}(x)\, y_2 + a_{23}(x)\, y_3 \\ y_3' = a_{31}(x)\, y_1 + a_{32}(x)\, y_2 + a_{33}(x)\, y_3 \end{array}\right\} \tag{147}$$

and introduce the *coefficient matrix*

$$A(x) = \begin{pmatrix} a_{11}(x) & a_{12}(x) & a_{13}(x) \\ a_{21}(x) & a_{22}(x) & a_{23}(x) \\ a_{31}(x) & a_{32}(x) & a_{33}(x) \end{pmatrix}$$

and the vector solution

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

which is a one-column matrix. For our further aims it should be noted that if we are given a matrix $B(x) = (b_{ij}(x))$ the rules of performing linear operations on matrices (see the beginning of Sec. XI.2) imply that $\frac{\Delta B}{\Delta x} = \left(\frac{\Delta b_{ij}}{\Delta x}\right)$ and hence $B'(x) = (b'_{ij}(x))$. This means that to differentiate a matrix we must differentiate all its elements. It is also easy to verify that all the basic rules of differentiation (such as the formulas for the derivative of a sum or of a product) remain true for matrices. It follows that

$$\mathbf{y}' = \begin{pmatrix} y'_1 \\ y'_2 \\ y'_3 \end{pmatrix}$$

and therefore, after a manner of (XI.9), system (147) can be written in the *matrix form*

$$\mathbf{y}' = A(x)\,\mathbf{y} \tag{148}$$

Accordingly, non-homogeneous linear system (140) can be put down in an analogous form

$$\mathbf{y}' = A(x)\,\mathbf{y} + \mathbf{f}(x), \quad \left(\mathbf{f}(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \end{pmatrix}\right)$$

Solution (139) can be rewritten in the vector form

$$\mathbf{y} = C_1\mathbf{y}^1 + C_2\mathbf{y}^2 + C_3\mathbf{y}^3$$

where the indices designate the numbers of the corresponding linearly independent particular vector solutions of equation (148). System (141) turns into

$$\mathbf{y}' = A\mathbf{y} \tag{149}$$

and its solution (142) is of the form

$$\mathbf{y} = e^{px}\boldsymbol{\alpha} \tag{150}$$

where

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

is a constant vector. Substituting (150) into (149) we get

$$pe^{px}\boldsymbol{\alpha} = Ae^{px}\boldsymbol{\alpha}, \quad \text{that is} \quad A\boldsymbol{\alpha} = p\boldsymbol{\alpha}$$

We see (compare with Sec. XI.4) that α and p must be, respectively, an eigenvector and the corresponding eigenvalue of the matrix A. As it was shown in Sec. XI.4, an eigenvalue is found from the equation

$$\det (A - pI) = 0$$

which is nothing but equation (144).

In solving systems of form (141) with constant coefficients (and also the corresponding non-homogeneous systems) we can apply the operator method (see Sec. 20). To do this we write $y' = Dy$, $z' = Dz$ and $u' = Du$ and then solve the resulting system of equations as an algebraic system in the unknowns y. z and u according to the methods of Sec. VI.4. But when doing this we stop performing algebraic operations when we arrive at formulas of type (VI.13) which are of the form $P(D) y = f(x)$ in our case. Then we use the methods of Sec. 20 and thus complete the process of solving the system. The procedure described here is nothing but another form of the method of reducing a first-order system to one equation of higher order.

22. Applications to Testing Lyapunov Stability of Equilibrium State. We understand the stability of an object, a state or a process as its ability to oppose external actions. It is often impossible to take some of these external factors into account beforehand. The concept of stability emerged as early as antiquity and now it plays an important role in physics and engineering. There are different concrete realizations of this general notion depending on the type of the object under consideration. on the character of external effects and so on. One of the realizations was discussed in our course in Sec. 16. Here we are going to consider the notion of the **Lyapunov stability** which is one of the most important forms of stability. It was introduced by the prominent Russian mathematician A. M. Lyapunov (1857-1918) in 1892.

Let the state of an object be described by means of a finite number of parameters. For definiteness, suppose that there are three such parameters x, y, z. Then a process in which the object changes as time passes is determined by three functions $x = x(t)$, $y = y(t)$, $z = z(t)$ (where t is time). Let the law of the process be described by a system of differential equations of the form

$$\left. \begin{array}{l} \dfrac{dx}{dt} = P(x. \, y, \, z) \\[2mm] \dfrac{dy}{dt} = Q(x. \, y, \, z) \\[2mm] \dfrac{dz}{dt} = R(x. \, y, \, z) \end{array} \right\} \qquad (151)$$

where the right-hand sides which are regarded as being known do not contain the independent variable t explicitly. The last condition means that the character of the differential law of development of the process does not change as time passes.

Suppose that an **equilibrium state** of the object, that is a state in which it does not change in time, is described by certain constant values $x = x_0$, $y = y_0$, $z = z_0$. Then these constants regarded as functions of the time must satisfy system (151). The direct substitution of x_0, y_0, z_0 into (151) implies that for this to be so it is necessary and sufficient that the relations

$$P(x_0, y_0, z_0) = 0, \quad Q(x_0, y_0, z_0) = 0,$$
$$R(x_0, y_0, z_0) = 0 \tag{152}$$

should hold simultaneously.

Suppose that at an instant t_0 the object is shifted from the equilibrium position, that is the parameters become equal to $x = x_0 + \Delta x_0$, $y = y_0 + \Delta y_0$ and $z = z_0 + \Delta z_0$. To investigate the further changes of the state of the object we must solve system (151) with the initial conditions

$$x(t_0) = x_0 + \Delta x_0, \quad y(t_0) = y_0 + \Delta y_0,$$
$$z(t_0) = z_0 + \Delta z_0 \tag{153}$$

The equilibrium state in question is said to be **stable (Lyapunov stable)** if after any infinitesimal displacement from the state the object remains all the time in an infinitesimal vicinity of the equilibrium state. In other words, the differences

$$\Delta x = x(t) - x_0, \quad \Delta y = y(t) - y_0, \quad \Delta z = z(t) - z_0$$

corresponding to the solutions of system (151) with initial conditions (153), for infinitesimal Δx_0, Δy_0, Δz_0, must be infinitesimal over the whole time interval $t_0 \leqslant t < \infty$.

At first sight one may find it strange that we consider infinitesimal deviations of the parameters and an infinite time interval because, practically, all the quantities are finite in reality. But here it is advisable to recall the difference between the mathematical and practical infinities (see Sec. III.1). A practical infinitesimal quantity is a real quantity which is small relative to the scale of the process in question. Similarly, a practically infinite time interval is the interval of a *transient process*, that is a process of passing from the state under consideration to a state of a different type (for instance, the transition from a state of equilibrium to another state of this type or from a state of equilibrium to the collapse of the object and so on). Hence, in reality the Lyapunov stability means that any small deviation from the equilibrium state does not practically change the state.   :

To test whether there is stability we substitute $x = x_0 + \Delta x$, $y = y_0 + \Delta y$ and $z = z_0 + \Delta z$ into system (151) which results in

$$\left.\begin{aligned}
\frac{d(\Delta x)}{dt} &= P(x_0 + \Delta x, \ y_0 + \Delta y, \ z_0 + \Delta z) = \\
&= (P'_x)_0 \Delta x + (P'_y)_0 \Delta y + (P'_z)_0 \Delta z + \ldots \\
\frac{d(\Delta y)}{dt} &= Q(x_0 + \Delta x, \ y_0 + \Delta y, \ z_0 + \Delta z) = \\
&= (Q'_x)_0 \Delta x + (Q'_y)_0 \Delta y + (Q'_z)_0 \Delta z + \ldots \\
\frac{d(\Delta z)}{dt} &= R(x_0 + \Delta x, \ y_0 + \Delta y, \ z_0 + \Delta z) = \\
&= (R'_x)_0 \Delta x + (R'_y)_0 \Delta y + (R'_z)_0 \Delta z + \ldots
\end{aligned}\right\} \quad (154)$$

where the notation $(P'_x)_0 = P'_x(x_0, y_0, z_0)$ etc. is introduced. In transforming the right-hand sides we have used Taylor's formula (XII.17) and formulas (152). Here the dots designate the terms of the order of smallness higher than the first.

When investigating the stability we consider only small values of Δx, Δy, Δz and therefore the most important role in the right-hand sides of system (154) belongs to the linear terms that are put down there. Therefore we replace system (154) by a *truncated system* (a *system of the first approximation*) which is obtained by dropping the terms of higher order of smallness and which has the form

$$\left.\begin{aligned}
\frac{d(\Delta x)}{dt} &= (P'_x)_0 \Delta x + (P'_y)_0 \Delta y + (P'_z)_0 \Delta z \\
\frac{d(\Delta y)}{dt} &= (Q'_x)_0 \Delta x + (Q'_y)_0 \Delta y + (Q'_z)_0 \Delta z \\
\frac{d(\Delta z)}{dt} &= (R'_x)_0 \Delta x + (R'_y)_0 \Delta y + (R'_z)_0 \Delta z
\end{aligned}\right\} \quad (155)$$

System (155) is a linear system with constant coefficients which can be solved by the method of Sec. 21. According to formula (145) (in which, of course, we had different notation) the general solution of system (155) is a linear combination of functions of the form e^{pt} where p satisfies the characteristic equation

$$\begin{vmatrix}
(P'_x)_0 - p & (P'_y)_0 & (P'_z)_0 \\
(Q'_x)_0 & (Q'_y)_0 - p & (Q'_z)_0 \\
(R'_x)_0 & (R'_y)_0 & (R'_z)_0 - p
\end{vmatrix} = 0 \quad (156)$$

To small Δx_0, Δy_0, Δz_0 there correspond small values of the arbitrary constants C_1, C_2, C_3. Therefore the behaviour of a solution, as $t \to \infty$, is completely determined by the behaviour of the functions e^{pt}. If $p = r + is$ (the case $s = 0$ is not excluded here) we have

$|e^{pt}| = e^{rt}$ [see formula (VIII.13)] and hence

$$|e^{pt}|_{t\to\infty} \to 0 \text{ for } r < 0 \quad \text{and} \quad |e^{pt}|_{t\to\infty} \to \infty \text{ for } r > 0 \quad (157)$$

Consequently, we arrive at the following conclusions: if all the roots of characteristic equation (156) have negative real parts (in particular, they may be negative real numbers) the state of equilibrium x_0, y_0, z_0 is stable. Besides, in this case we have $x(t) \to x_0$, $y(t) \to y_0$ and $z(t) \to z_0$ for small Δx_0, Δy_0, Δz_0, as $t \to \infty$. In such a case we say that the equilibrium state is **asymptotically stable**. But if there is at least one root with a positive real part the state of equilibrium is unstable.

We have derived these results from system (155) but according to the above considerations the same assertions are true for the complete system, i.e. system (154). It should be noted that our conclusions also remain true for the case of multiple roots of equation (156) although then we can have powers of t as factors entering into the solution. In fact, exponential function (157) approaching zero, for $r < 0$, faster than any negative power of t, the above assertion appears evident.

Both conclusions obtained above do not include the case when there are no roots with positive real parts among the roots of equation (156) but there is at least one root having a zero real part. In this case there appear functions of the form

$$e^{ist} = \cos st + i \sin st \quad (|e^{ist}| = 1)$$

in the general solution of system (155). This implies that the object is likely to oscillate about the equilibrium position or to remain motionless near this position without approaching it. But the time interval being infinite, the terms of higher order of smallness that have been dropped begin to influence the process noticeably, and this can violate the stability. It is therefore impossible to arrive at any conclusions on the stability or instability of the state of equilibrium in this special case judging by the roots of equation (156). To investigate the stability in such a case we must involve some additional considerations; for instance, we can try to consider subsequent terms in expansions (154).

In the case when the changes in our object are described by means of one function $x(t)$ satisfying the differential equation

$$\frac{dx}{dt} = f(x) \tag{158}$$

the above results are especially simple. We see that if $f(x_0) = 0$ and $f'(x_0) < 0$ the value $x = x_0$ corresponds to a stable equilibrium state, and if $f(x_0) = 0$ and $f'(x_0) > 0$ the state is unstable. [We

suggest that the reader should draw this conclusion on the basis of the disposition of the isoclines of equation (158) in the t, x-plane.]

There are many books on the theory of stability. We refer the reader to a comprehensive course [31].

§ 7. Approximate and Numerical Methods of Solving Differential Equations

We often cannot integrate a differential equation exactly by reducing it to quadratures. Then we should apply other methods for constructing a solution. We have already described the simplest graphical method for solving first-order equations (see Sec. 3). Here we are going to present some methods for constructing approximate formulas of solutions which are analogous to the methods of solving finite equations described in § V.1. We shall also discuss some numerical methods of solving differential equations which yield a sought-for particular solution represented in a tabular form. For the sake of simplicity, we shall consider first-order equations but the same methods can naturally be extended to equations of an arbitrary order and to systems of equations.

23. Iterative Method. Let us take a first-order differential equation with a given initial condition of the form

$$y' = f(x, y), \quad y(x_0) = y_0 \tag{159}$$

Taking integrals of both sides of the equation we obtain

$$\int_{x_0}^{x} y' \, dx = y - y_0 = \int_{x_0}^{x} f(x, y) \, dx = \int_{x_0}^{x} f(x, y(x)) \, dx$$

Changing the notation for the variable of integration we write

$$y(x) = y_0 + \int_{x_0}^{x} f(s, y(s)) \, ds \tag{160}$$

Equation (160) is equivalent to both equalities (159) because if we differentiate it we obtain the first equality and if we substitute $x = x_0$ in it we obtain the second one. Equation (160) is an **integral equation** since the unknown function appears under the sign of integration in the equation.

The form of equation (160) is convenient for applying the iterative method to it [compare with equation (V.9)] although here we have an unknown quantity that is a function but not a number. Now we must choose a certain function $y_0(x)$ as a zero approximation (initial approximation). It is desirable to choose it so that it should be as close as possible to the sought-for solution. If we have no information about the solution we can simply put $y_0(x) \equiv y_0$. The

first approximation is then found by the formula

$$y_1(x) = y_0 + \int\limits_{x_0}^{x} f(s, y_0(s))\, ds$$

Substituting the result into the right-hand side of (160) we obtain the second approximation and so on. Generally,

$$y_{n+1}(x) = y_0 + \int\limits_{x_0}^{x} f(s, y_n(s))\, ds \quad (n = 0, 1, 2, \ldots) \tag{161}$$

As in Sec. V.3, if the iterative process converges. that is if the successive approximations tend to a certain limiting function, as n increases, this function satisfies equation (160). To verify this we can pass to the limit in equality (161) as $n \to \infty$.

It is remarkable that the iterative method applied to equation (160) usually converges for all x which are close enough to x_0. At any rate. this is so if the conditions of Cauchy's theorem (see Sec. 3) are fulfilled. The convergence is due to the fact that in calculating subsequent approximations we integrate the preceding ones and that successive integrations "smooth" the function and gradually eliminate various irregularities which have been brought in by the choice of the zero approximation, by the round-off errors and the like. In contrast to it. differentiation of a function, as a rule, "worsens" the function and increases the initial irregularities. Therefore an iterative method based on successive differentiations is likely to diverge.

The difference between integration and differentiation is illustrated in Fig. 298. There is a disturbance of the function depicted in the upper part of Fig. 298. We see

Fig. 298

that this essentially changes the derivative of the function which is shown below (what is the form of the graph of the second derivative?). At the same time we see that the integral is changed very slightly.

For example, let us consider a particular case of Riccati's equation (see the end of Sec. 4) of the form

$$y' = x^2 + y^2$$
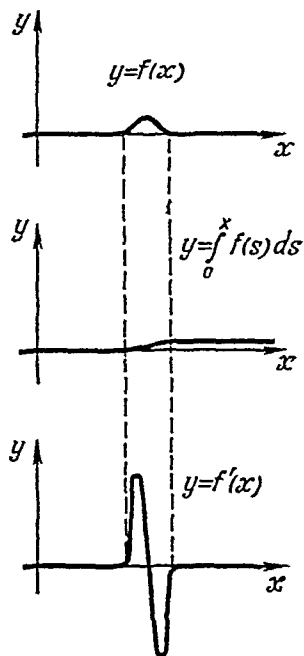
with the initial condition $y(0) = 0$. After integrating we have

$$y(x) = \frac{x^3}{3} + \int\limits_0^x y^2(s)\, ds$$

We have no information about the desired solution and therefore we choose the zero function $y_0(x) \equiv 0$ as a zero approximation which, at any rate, satisfies the initial condition. Then we obtain (verify the calculations!)

$$y_1(x) = \frac{x^3}{3}, \quad y_2(x) = \frac{x^3}{3} + \int\limits_0^x \left(\frac{s^3}{3}\right)^2 dx = \frac{x^3}{3} + \frac{x^7}{63},$$

$$y_3(x) = \frac{x^3}{3} + \frac{x^7}{63} + \frac{2x^{11}}{2079} + \frac{x^{15}}{59535}$$

and so on. We see that for small x (for instance, for $|x| < 1$) the process converges fast. Really, we can put $y = \frac{x^3}{3} + \frac{x^7}{63}$ for $|x| < 1$ with an accuracy of 0.001, and for $|x| < \frac{1}{2}$ the same accuracy is attained if we simply set $y = \frac{x^3}{3}$.

The question as to what is the approximation at which we should stop the iterative process is answered by comparing subsequent approximations with the preceding ones.

24. Application of Taylor's Series. Differentiating equation (159) and using the initial condition we can find the values $y'(x_0)$, $y''(x_0)$ etc. Therefore we can form an expansion of the solution into Taylor's series (see Sec. IV.16). The necessary number of terms that guarantees a chosen degree of accuracy is found by successive calculation of the terms and their comparison with the degree of accuracy.

For instance, let us take the problem

$$y' = x^2 + y^2, \quad y(0) = 1$$

Substituting $x = 0$ into the right-hand side of the equation we find

$$y'(0) = 0^2 + 1^2 = 1$$

Differentiating both sides of the equation we obtain $y'' = 2x + 2yy'$ and then, substituting $x = 0$, we derive

$$y''(0) = 2\cdot 0 + 2\cdot 1\cdot 1 = 2$$

We likewise find

$$y''' = 2 + 2y'^2 + 2yy'', \quad y'''(0) = 8, \quad y^{\mathrm{IV}} = 6y'y'' + 2yy''',$$
$$y^{\mathrm{IV}}(0) = 28$$

and so on. Substituting the results into Maclaurin's formula (IV.54) we obtain

$$y = y(0) + \frac{y'(0)}{1!}x + \frac{y''(0)}{2!}x^2 + \ldots =$$

$$= 1 + x + x^2 + \frac{4}{3}x^3 + \frac{7}{6}x^4 + \ldots$$

The formula can be used for small values of $|x|$.

25. Application of Power Series with Undetermined Coefficients. This method is closely related to Sec. 24. According to the method we seek a solution of a given equation in the form of a series with unknown coefficients of the form

$$y = a + b(x - x_0) + c(x - x_0)^2 + d(x - x_0)^3 + \ldots \quad (162)$$

The coefficients are found after we substitute the series into the equation and equate the coefficients of the same powers of x (and use the initial conditions if there are any).

For example, let us consider Airy's equation

$$y'' - xy = 0 \quad (163)$$

We shall look for a solution in the form of an expansion into powers of x:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots \quad (164)$$

After (164) is differentiated and substituted into the equation we obtain

$$(1 \cdot 2a_2 + 2 \cdot 3a_3 x + 3 \cdot 4a_4 x^2 + \ldots) - x(a_0 + a_1 x + a_2 x^2 + \ldots) = 0$$

Equating coefficients in equal powers of x we derive

$$1 \cdot 2a_2 = 0, \quad 2 \cdot 3a_3 - a_0 = 0, \quad 3 \cdot 4a_4 - a_1 = 0,$$
$$4 \cdot 5a_5 - a_2 = 0, \quad 5 \cdot 6a_6 - a_3 = 0, \quad \ldots$$

from which we find, in succession,

$$a_2 = 0, \quad a_3 = \frac{a_0}{2 \cdot 3}, \quad a_4 = \frac{a_1}{3 \cdot 4}, \quad a_5 = \frac{a_2}{4 \cdot 5} = 0,$$

$$a_6 = \frac{a_3}{5 \cdot 6} = \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6}, \quad a_7 = \frac{a_4}{6 \cdot 7} = \frac{a_0}{3 \cdot 4 \cdot 6 \cdot 7}, \quad a_8 = \frac{a_5}{7 \cdot 8} = 0,$$

$$a_9 = \frac{a_6}{8 \cdot 9} = \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6 \cdot 8 \cdot 9}$$

and so on.

The substitution of these results into formula (164) yields the general solution of equation (163):

$$y = a_0 + a_1 x + \frac{a_0}{2 \cdot 3} x^3 + \frac{a_1}{3 \cdot 4} x^4 + \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6} x^6 +$$

$$+ \frac{a_1}{3 \cdot 4 \cdot 6 \cdot 7} x^7 + \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6 \cdot 8 \cdot 9} x^9 + \ldots =$$

$$= a_0 \left(1 + \frac{x^3}{2 \cdot 3} + \frac{x^6}{2 \cdot 3 \cdot 5 \cdot 6} + \frac{x^9}{2 \cdot 3 \cdot 5 \cdot 6 \cdot 8 \cdot 9} + \cdots \right) +$$

$$+ a_1 \left(x + \frac{x^4}{3 \cdot 4} + \frac{x^7}{3 \cdot 4 \cdot 6 \cdot 7} + \frac{x^{10}}{3 \cdot 4 \cdot 6 \cdot 7 \cdot 9 \cdot 10} + \cdots \right)$$

The constants a_0 and a_1 are left undetermined and play the role of arbitrary constants which were previously designated by C_1 and C_2 in Sec. 14 (see property 5 in the section). The series which are taken inside the parentheses are two linearly independent particular solutions of equation (163). In particular, putting $a_0 = \left[\sqrt[3]{9}\, \Gamma \left(\frac{2}{3} \right) \right]^{-1}$ and $a_1 = - \left[\sqrt[3]{3}\, \Gamma \left(\frac{1}{3} \right) \right]^{-1}$ we obtain the so-called *Airy's function of the first kind* which is denoted as $Ai(x)$.

The above technique is always applicable to linear equations of the form

$$a_0(x) y^{(n)} + a_1(x) y^{(n-1)} + \ldots + a_n(x) y = f(x) \qquad (165)$$

in case all the functions $a_0(x)$, $a_1(x)$, ..., $f(x)$ are polynomials in x or, in a more general case, when they are sums of power series in powers of $x - x_0$, and $a_0(x_0) \neq 0$.

If $a_0(x_0) = 0$ the value x_0 is said to be a **singular point** of equation (165). Then there may be no solution of form (162). In this case it is sometimes possible to find a solution in the form

$$y = (x - x_0)^\rho [a + b(x - x_0) + c(x - x_0)^2 +$$
$$+ d(x - x_0)^3 + \ldots] \qquad (166)$$

where the constant ρ should also be chosen. When selecting a solution of this form we can regard a as being unequal to zero because otherwise we can take a certain power of $x - x_0$ outside the brackets and thus the operation reduces to a change of ρ in its value.

26. **Bessel's Functions.** We now consider an important example of the so-called *Bessel's equation* of the form

$$x^2 y'' + xy' + (x^2 - p^2) y = 0 \qquad (167)$$

where $p = \text{const} \geqslant 0$ and $0 < x < \infty$. Solutions of the equation are called **Bessel's functions** although in fact they were used by Euler beginning with 1766, that is before Bessel was born. The

functions are also called cylindrical functions because they are widely applied to solving equations of mathematical physics in a domain of the form of a circular cylinder.

Since the point $x = 0$ is a singular point of equation (167) its solution should be sought for according to formula (166) in which we must set $x_0 = 0$:

$$y = ax^\rho + bx^{\rho+1} + cx^{\rho+2} + dx^{\rho+3} +$$
$$+ ex^{\rho+4} + fx^{\rho+5} + gx^{\rho+6} + ix^{\rho+7} + jx^{\rho+8} + \ldots \quad (168)$$

The differentiation of (168) and the substitution into (167) result in

$$x^2 [a\rho(\rho-1) x^{\rho-2} + b(\rho+1) \rho x^{\rho-1} + c(\rho+2)(\rho+1) x^\rho + \ldots] +$$
$$+ x [a\rho x^{\rho-1} + b(\rho+1) x^\rho + c(\rho+2) x^{\rho+1} + \ldots] +$$
$$+ x^2 (ax^\rho + bx^{\rho+1} + cx^{\rho+2} + \ldots) - p^2 (ax^\rho + bx^{\rho+1} + cx^{\rho+2} + \ldots) = 0$$

After the coefficients in the same powers of x are equalled we obtain an infinite number of equalities which are then solved in succession: |

$$a\rho(\rho-1) + a\rho - ap^2 = 0 \text{ yields } a(\rho^2 - p^2) = 0,$$
$$b(\rho+1)\rho + b(\rho+1) - bp^2 = 0 \text{ yields } b(\rho^2 + 2\rho + 1 - p^2) = 0,$$
$$c(\rho+2)(\rho+1) + c(\rho+2) + a - cp^2 = 0 \text{ yields }$$
$$c(\rho^2 + 4\rho + 4 - p^2) + a = 0,$$
$$d(\rho+3)(\rho+2) + d(\rho+3) + b - dp^2 = 0 \text{ yields }$$
$$d(\rho^2 + 6\rho + 9 - p^2) + b = 0,$$
$$e(\rho+4)(\rho+3) + e(\rho+4) + c - ep^2 = 0 \text{ yields }$$
$$e(\rho^2 + 8\rho + 16 - p^2) + c = 0$$

and so on. Since $a \neq 0$ we see that, by the first equality, we have $\rho^2 = p^2$, i.e. $\rho = \pm p$. Substituting the result into the remaining equalities we receive, in succession,

$$b = 0, \quad c = \frac{-a}{4\rho+4} = -\frac{a}{2^2(\rho+1)}, \quad d = 0,$$

$$e = -\frac{c}{8\rho+16} = \frac{a}{2^4 \cdot 2(\rho+1)(\rho+2)}, \quad f = 0,$$

$$g = -\frac{a}{2^6 \cdot 2 \cdot 3(\rho+1)(\rho+2)(\rho+3)}, \quad i = 0,$$

$$j = \frac{a}{2^8 \cdot 2 \cdot 3 \cdot 4(\rho+1)(\rho+2)(\rho+3)(\rho+4)} \text{ etc.}$$

From this, on ¦the basis of formula (168), we find a solution
of the form

$$y = ax^\rho - \frac{a}{2^2(\rho+1)} x^{\rho+2} + \frac{a}{2^4 \cdot 2!(\rho+1)(\rho+2)} x^{\rho+4} -$$

$$- \frac{a}{2^6 \cdot 3!(\rho+1)(\rho+2)(\rho+3)} x^{\rho+6} + \ldots \qquad (169)$$

where a is an arbitrary constant. Here it is convenient to put
$a = \dfrac{1}{2^\rho \Gamma(\rho+1)}$ (see Sec. XIV.17). By formula (XIV.66), we have

$$\Gamma(\rho+1)(\rho+1) = \Gamma(\rho+2), \quad \Gamma(\rho+1)(\rho+1)(\rho+2) =$$
$$= \Gamma(\rho+2)(\rho+2) = \Gamma(\rho+3) \text{ etc.}$$

and therefore formula (169) implies, for the above a, that

$$y = \frac{1}{\Gamma(\rho+1)}\left(\frac{x}{2}\right)^\rho - \frac{1}{1!\Gamma(\rho+2)}\left(\frac{x}{2}\right)^{\rho+2} +$$

$$+ \frac{1}{2!\Gamma(\rho+3)}\left(\frac{x}{2}\right)^{\rho+4} - \frac{1}{3!\Gamma(\rho+4)}\left(\frac{x}{2}\right)^{\rho+6} + \ldots =$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\Gamma(\rho+n+1)}\left(\frac{x}{2}\right)^{\rho-2n} \qquad (170)$$

This sum is called Bessel's function of the first kind of order ρ
and is denoted as $J_\rho(x)$. Since $\rho = \pm p$ the general solution of equa-
tion (167) (see property 5 in Sec. 14) can be written as

$$y = C_1 J_p(x) + C_2 J_{-p}(x) \qquad (171)$$

Formula of the general solution (171) does not apply for integer
$p = 0, 1, 2, 3, \ldots$. Actually, for such p and $\rho = -p$ we have

$$\Gamma(-p-1) = \Gamma(-p+2) = \ldots = \Gamma(-p+p) = \pm \infty$$

and therefore formula (170) results in

$$J_{-p}(x) = \sum_{n=p}^{\infty} \frac{(-1)^n}{n!(-p+n)!}\left(\frac{x}{2}\right)^{-p+2n} =$$

$$= \sum_{n'=0}^{\infty} \frac{(-1)^p(-1)^{n'}}{(p+n')!(n')!}\left(\frac{x}{2}\right)^{p+2n'} = (-1)^p J_p(x) \quad (p = 0, 1, 2, \ldots)$$

(we have made the substitution $n - p = n'$). Consequently, in
this case the solutions $J_p(x)$ and $J_{-p}(x)$ are linearly dependent
(see Sec. 14), and formula (171) does not therefore express the general
solution (although it yields a family of particular solutions depen-
ding on one essential parameter).

To obtain a formula for the general solution of equation (167) valid for all p we can apply the operation similar to the one used in Sec. 17 (see case 3). Namely, we first suppose that p is non-integral, and form the function

$$Y_p(x) = \cot p\pi J_p(x) - \frac{1}{\sin p\pi} J_{-p}(x) = \frac{\cos p\pi J_p(x) - J_{-p}(x)}{\sin p\pi}$$

The last function being a linear combination of the solutions, we obtain a solution of equation (167) which is called **Bessel's function of the second kind of order p**. It is also sometimes designated as $N_p(x)$. Now, if p becomes an integer, there appears an indeterminate form on the right-hand side (why?). It can be calculated according to L'Hospital's rule but we shall not put down the calculations here. We only note that the result will be a sum in which the expression

$$-\frac{(p-1)!\, 2^p}{\pi x^p} \quad \text{(for } p = 1, 2, 3, \ldots) \quad \text{or} \quad \frac{2}{\pi} \ln x$$

$$\text{(for } p = 0)$$

will be the principal term as $x \to 0$.

Thus, the formula

$$y = C_1 J_p(x) + C_2 Y_p(x) \tag{172}$$

represents the general solution of equation (167) for all $p \geqslant 0$ (both integral and non-integral) in the interval $0 < x < \infty$. The value $J_p(+0)$ is finite here whereas $Y_p(+0) = -\infty$. Therefore if the conditions of a problem imply that $y(+0)$ should be finite we must retain only the first summand on the right-hand side of formula (172).

Bessel's functions have been thoroughly studied, and there exist extensive tables for the functions. The functions

$$\left.\begin{aligned} J_0(x) &= 1 - \frac{x^2}{(1!)^2\, 2^2} + \frac{x^4}{(2!)^2\, 2^4} - \frac{x^6}{(3!)^2\, 2^6} + \cdots \\ J_1(x) &= \frac{x}{2} - \frac{x^3}{1!2!2^3} + \frac{x^5}{2!3!2^5} - \frac{x^7}{3!4!2^7} + \cdots \end{aligned}\right\} \tag{173}$$

are the most important for applications. The graphs of these functions are approximately represented in Fig. 299. These functions and all Bessel's functions of the first and of the second kind change their signs infinitely many times and tend to zero as x increases. From formulas (173) we can easily deduce the relationship $J_0'(x) = -J_1(x)$ (check it up!). There are some other relationships between Bessel's functions. All these properties can be found in [29].

27. Small Parameter Method. This method was described in Sec. V.5. It can also be applied to solving differential equations. Here we present some simple examples.

The problem

$$y' = \frac{x}{1+0.1xy}, \quad y(0) = 0 \tag{174}$$

does not involve any parameters. But we can take a more general problem of the form

$$y' = \frac{x}{1+\alpha xy}, \quad y(0) = 0 \tag{175}$$

where α is a small parameter. Problem (174) is a particular case of (175) for $\alpha = 0.1$. Problem (175) can be easily solved for $\alpha = 0$.
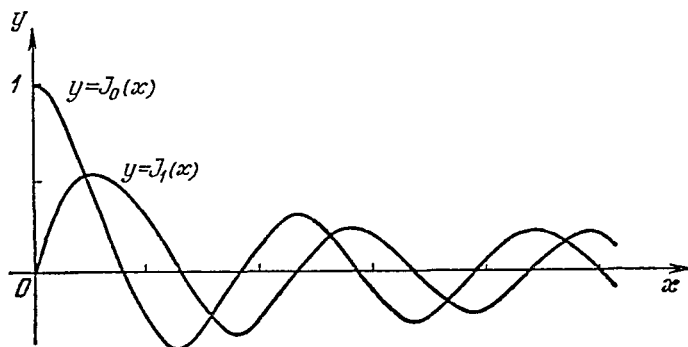
Fig. 299

Evidently, in this case we have $y = \frac{x^2}{2}$. Therefore we seek the solution of problem (175) as an expansion in powers of α, that is in the form

$$y = \frac{x^2}{2} + \alpha u + \alpha^2 v + \alpha^3 w + \ldots \tag{176}$$

where the functions $u = u(x)$, $v = v(x)$ etc. depending on x are yet unknown.

Substituting (176) into (175) and multiplying by the denominator we obtain

$$(x + \alpha u' + \alpha^2 v' + \alpha^3 w' + \ldots)\left(1 + \frac{\alpha}{2}x^3 + \alpha^2 xu + \alpha^3 xv + \ldots\right) = x \tag{177}$$

The initial condition implies

$$\alpha u(0) + \alpha^2 v(0) + \ldots = 0$$

and therefore we have

$$u(0) = 0. \quad v(0) = 0, \quad w(0) = 0, \ldots \tag{178}$$

Removing the parentheses in (177) and equating the coefficients in power of α to zero we get, in succession,

$$u' + \frac{1}{2} x^4 = 0, \quad v' + \frac{x^3}{2} u' + x^2 u = 0,$$

$$w' + \frac{x^3}{2} v' + xuu' + x^2 v = 0$$

and so on. Now taking into account equalities (178) we find

$$u = -\frac{x^5}{10}, \quad v = \frac{7}{160} x^8, \quad w = \frac{71}{1760} x^{11}$$

etc. (check up the calculations!).

Consequently, formula (176) yields

$$y = \frac{x^2}{2} - \frac{\alpha}{10} x^5 + \frac{7\alpha^2}{160} x^8 - \frac{71\alpha^3}{1760} x^{11} + \cdots$$

In particular. putting $\alpha = 0.1$. we obtain the expression

$$y = \frac{x^2}{2} - \frac{x^5}{100} + \frac{7x^8}{16,000} - \frac{71x^{11}}{1,760,000} + \cdots$$

for the solution of equation (174). This series perfectly converges for $|x| < 1$. the convergence being a little slower for $1 < |x| < 2$.

As another example, let us take the problem

$$y' = \sin(xy), \quad y(0) = \alpha \tag{179}$$

Here, in |contrast to the previous problem, there is a parameter entering into the initial condition. Problem (179) has an apparent solution for $\alpha = 0$, namely the solution $y \equiv 0$. Therefore we look for the solution of the form

$$y = \alpha u + \alpha^2 v + \alpha^3 w + \cdots \quad (u = u(x), \ v = v(x), \ \ldots) \tag{180}$$

for small $|\alpha|$. The substitution of the value $x = 0$ yields

$$u(0) = 1, \quad v(0) = 0, \quad w(0) = 0, \ \ldots \tag{181}$$

On the other hand, substituting (180) into differential equation (179) and taking into account the power series for the sine [see formula (IV.57)] we receive

$$\alpha u' + \alpha^2 v' + \alpha^3 w' + \cdots =$$
$$= \frac{(\alpha xu + \alpha^2 xv + \alpha^3 xw + \cdots)}{1!} - \frac{(\alpha xu + \alpha^2 xv + \alpha^3 xw + \cdots)^3}{3!} + \cdots$$

Equating the coefficients of the same powers of α we get

$$u' = xu, \quad v' = xv, \quad w' = xw - \frac{x^3 u^3}{3!}, \ \ldots$$

Hence, here we have arrived at linear equations which must be solved in succession. Integrating the equations with initial con-

ditions (181) we find

$$u = e^{\frac{x^2}{2}}, \quad v = 0, \quad w = \frac{1}{12}(1-x^2)e^{\frac{3}{2}x^2} - \frac{1}{2}e^{\frac{x^2}{2}}, \ldots$$

(check up the results!). Substituting the above expressions into (180) we obtain the sought-for solution in the form of an expansion which is valid for small $|x|$ and $|\alpha|$.

In more complicated problems involving the small parameter method it often turns out that even the determination of the first term containing the parameter may yield fruitful results.

28. General Remarks on Dependence of Solutions on Parameters. Here we give some general considerations related to the problems discussed in the preceding section. There are many cases when a differential equation in question or a system of such equations contains one or more parameters which can take on different constant values. For simplicity's sake, let us consider a first-order equation of the form

$$\frac{dy}{dx} = f(x, y, \lambda) \tag{182}$$

where λ is a parameter. Suppose we have certain initial conditions $x = x_0$, $y = y_0$.

Let us suppose that the point (x_0, y_0) is not a singular point (see Sec. 7). This means that there is a single solution of equation (182) satisfying the initial conditions. Then the geometric meaning of equation (182) (see Sec. 3) implies that if its right-hand side is continuous in λ the change of the direction field corresponding to small variations of λ will also be small. Therefore the solution $y(x, \lambda)$ will also be continuous in λ. We can likewise conclude that the situation will be the same if not only the right-hand side of the equation continuously depends on λ but the initial conditions as well, that is if $x_0 = x_0(\lambda)$ and $y_0 = y_0(\lambda)$.

Suppose that the solution $y(x, \lambda)$ of equation (182) is known for a certain value λ (the "non-perturbed" value of the parameter, as it is called). Now, let the value of the parameter change and become equal to $\lambda + \Delta\lambda$ where $|\Delta\lambda|$ is small. Then y will also change and gain an increment $\Delta_\lambda y$. The differential of $\Delta_\lambda y$, that is its principal linear part, will be called a **variation** of the solution and will be designated as δy.

Thus, a variation is a particular differential of a solution corresponding to an increment of the parameter. The new notation has been introduced to distinguish between the differentials corresponding to the argument x and to the parameter λ. When it is permissible to neglect infinitesimals of higher order of smallness we can simply say that the variation of a solution is an infinitesimal change of the solution due to an infinitesimal change of the parameter. With a

value of λ chosen, the quantity δy (as well as y) depends on x. Hence we can write $\delta y = \delta y\,(x)$ where $\delta y\,(x)$ also depends on $\Delta\lambda$ and is directly proportional to it (although the dependence on $\Delta\lambda$ is not indicated explicitly here).

To form a differential equation for δy we must equate the differentials (taken with respect to λ) of both sides of equality (182):

$$\delta\,\frac{dy}{dx} = \delta\,(f\,(x,\,y,\,\lambda)) = \frac{\partial f}{\partial y}\,\delta y + \frac{\partial f}{\partial \lambda}\,\delta\lambda \qquad (\delta\lambda = \Delta\lambda)$$

that is

$$\frac{d\,(\delta y)}{dx} = f'_y\,(x,\,y,\,\lambda)\,\delta y + f'_\lambda\,(x,\,y,\,\lambda)\,\delta\lambda \qquad (183)$$

In deducing (183) we have interchanged the signs d and δ because these are the signs of the differentials taken with respect to different variables (see Sec. IX.15). We have also applied the formula for differentiation of a composite function. Equation (183) is called the *variational equation* corresponding to original equation (182). Since the "non-perturbed" solution $y\,(x,\,\lambda)$ is substituted for y in the right-hand side of (183) the equation is linear in δy and can be easily integrated (see Sec. 4). Variational equations which correspond to higher-order equations or to systems of equations are not integrable by quadratures in the general case but they are always linear.

Let us establish the initial condition for δy. In the general case, when $x_0 = x_0\,(\lambda)$ and $y_0 = y_0\,(\lambda)$, we substitute $\lambda + \delta\lambda$ for λ and neglect infinitesimals of higher order which implies that we have $y = y_0\,(\lambda + \delta\lambda) = y_0 + y'_0\delta\lambda$ for $x = x_0\,(\lambda + \delta\lambda) = x_0 + x'_0\delta\lambda$ where $x'_0 = \frac{dx_0\,(\lambda)}{d\lambda}$ and $y'_0 = \frac{dy_0\,(\lambda)}{d\lambda}$. Hence, for the value $\lambda + \delta\lambda$ of the parameter and for $x = x_0$, we have

$$y\,|_{x=x_0} = y\,|_{x=x_0+x'_0\delta\lambda} - \partial_x y\,|_{dx=x'_0\delta\lambda} =$$

$$= y_0 + y'_0\delta\lambda - \frac{dy}{dx}\,x'_0\delta\lambda = y_0 + y'_0\delta\lambda - f_0 x'_0\delta\lambda$$

where $f_0 = f\,(x_0,\,y_0,\,\lambda)$. But the same value of y is equal to $(y\,(x,\,\lambda) + \delta y)\,|_{x=x_0} = y_0 + (\delta y)|_{x=x_0}$. Therefore, the initial value of δy is expressed by the following initial condition:

$$(\delta y)\,|_{x=x_0} = (y'_0 - f_0 x'_0)\,\delta\lambda$$

In the particular case of values x_0 and y_0 independent of λ we have $x'_0 = y'_0 = 0$ and therefore the initial condition has the form $(\delta y)|_{x=x_0} = 0$.

There are some cases when a parameter enters into a differential equation in such a way that the order of the equation reduces, that

is the equation degenerates, for certain values of the parameter. Such a situation is connected with new phenomena which we shall illustrate by taking an example.

Consider the problem

$$\lambda y' + y = 0, \qquad y\,|_{x=x_0} = 1 \tag{184}$$

whose solution is $y = e^{-\frac{x}{\lambda}}$. The value $\lambda = 0$ yields the degeneration (why?). Let the solution be considered for $x \geqslant 0$ and let $\lambda \to +0$; the solution is depicted in Fig. 300. The limiting case for equation (184), as $\lambda \to +0$, is the equation $y = 0$. This is a finite equation whose solution $y \equiv 0$ does not satisfy the initial condition $y|_{x=x_0} = 1$. Besides, we see that when λ is close to zero (but unequal to zero) the solution is close to zero for the values of x which are not too close to $x = 0$ (for instance, for the values of x exceeding the value h shown in Fig. 300), this not being the case for the values of x which are close to $x = 0$ (for instance, for the values of x lying in the interval $0 < x < h$). An interval of the type $0 < x < h$ (which we have in our particular problem) is called a *boundary layer*. The solution of (184) must cross the boundary layer $0 < x < h$ in order to pass from the unit initial value (184) to a value which is close to zero.
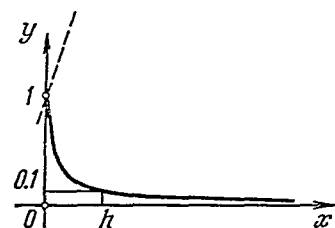
Fig. 300

The width of a boundary layer is understood conditionally. In fact, in our example the solution never turns into zero. If we conditionally take a certain value of x as the width of the boundary layer, for instance, $x = h$ for which the magnitude of the solution decreases 10 times in comparison with the initial magnitude, then, in the case of problem (184), we obtain

$$e^{-\frac{h}{\lambda}} = 0.1 \quad \text{and} \quad h = \ln 10 \cdot \lambda$$

that is the width of the boundary layer is directly proportional to the value of λ.

If $\lambda \to -0$ we obtain the solution which is shown in the dotted line in Fig. 300. This solution tends to infinity for any $x > 0$ as $\lambda \to -0$. This case is not as interesting as the preceding one.

An analogous phenomenon is often encountered in more complicated cases. For instance, let the solution of a second-order equation satisfying two initial conditions or boundary conditions (see Sec. 16) be considered and let the order of the equation be reduced by unity for a certain value $\lambda = \lambda_0$. Then the solution $y_0\,(x)$ corresponding

to $\lambda = \lambda_0$ satisfies a first-order equation. Therefore it often happens that if the solution is finite for $\lambda = \lambda_0$ it satisfies only one of the initial conditions and may not satisfy the other. The solution y (x) has a boundary layer for the values of λ close to λ_0 (with the width proportional to $|\lambda - \lambda_0|$) which is transversed by the solution when it passes from the other condition to y_0 (x). A similar situation may also occur for systems of differential equations.

29. Methods of Minimizing Discrepancy. The methods are based on the idea that an unknown function is sought in a form containing several parameters, that is in the form

$$y = \varphi (x, \lambda_1, \lambda_2, \ldots, \lambda_m) \tag{185}$$

The right-hand side is usually chosen in such a way that the initial or boundary conditions imposed on a solution should be satisfied for any values of the parameters. After expression (185) has been substituted into a given differential equation we obtain the corresponding discrepancy, that is the difference between the right-hand and left-hand sides. The discrepancy (which we denote by h) depends upon the parameters:

$$h = h (x, \lambda_1, \lambda_2, \ldots, \lambda_m)$$

If solution (185) were exact the discrepancy h would equal zero identically. Therefore we can impose m additional conditions that should be satisfied by the discrepancy and that are automatically fulfilled for the function which is identically equal to zero in order to determine the necessary values of the parameters $\lambda_1, \lambda_2, \ldots, \lambda_m$. For instance, we can equate h to zero for m different values of x; this is the **collocation method**. We can try to minimize the integral $\int_a^b h^2 \, dx$ on the interval $a \leqslant x \leqslant b$ in which the solution is being constructed; this is the **method of least squares**. We can equate to zero m integrals of the form $\int_a^b h\psi_1 (x) \, dx, \int_a^b h\psi_2(x) \, dx, \ldots$

$\ldots \int_a^b h\psi_m (x) \, dx$ where $\psi_1(x), \psi_2 (x), \ldots, \psi_m (x)$ is a chosen system of functions; this is the **method of moments** (integrals of this type are called *moments*).

The greater the number of the parameters we introduce, the more "flexible" formula (185) (that is the greater the accuracy of the representation of a sought-for solution which can be attained by means of the formula). But at the same time every increase of the number of the parameters leads to more and still more complicated calculations. It is an art to be able to forecast the form of the sought-for

solution by using a formula containing a few parameters. The correctness of the result can be judged by comparing the results of repeated calculations performed according to different methods or with the help of different numbers of parameters etc.

We can also take a right-hand side of formula (185) which identically satisfies only a number of initial or boundary conditions imposed on the solution. Then the corresponding number of conditions chosen for determining the values of the parameters must imply that the remaining conditions should be satisfied. It is apparent that in such a case the number of additional conditions that can be chosen more or less arbitrarily is respectively reduced.

Let us take an example in which it is possible to compare approximate solutions with the exact solution. Let it be necessary to solve the problem

$$y'' + y = 0 \quad (0 \leqslant x \leqslant 1), \quad y(0) = 0, \quad y(1) = 1$$

Let us look for the solution of the form

$$y = \lambda x + \mu x^2 \tag{186}$$

Here the first boundary condition is satisfied automatically whereas the second implies $\lambda + \mu = 1$. From this we obtain $y = \lambda x + (1 - \lambda) x^2$ and hence we have only one degree of freedom at our disposal, that is only one additional condition that we can choose to minimize the discrepancy which has the form

$$h = y'' + y = 2(1 - \lambda) + \lambda x + (1 - \lambda) x^2$$

in our case. The collocation method applied to $x = \frac{1}{2}$ yields the value $\lambda = \frac{9}{7}$. The method of least squares used for the interval $0 \leqslant x \leqslant 1$ yields the value $\lambda = \frac{257}{202}$. Finally, the method of moments with the function $\psi(x) \equiv 1$ yields the value $\lambda = \frac{14}{11}$ (let the reader verify all the calculations!). Substituting these values of λ into (186) we obtain the corresponding expressions which approximate the exact solution $y = \frac{\sin x}{\sin 1}$ fairly well. For example, the exact solution is equal to 0.5699 for $x = 0.5$ whereas the approximations yield, respectively, the values 0.5714, 0.5681 and 0.5682. Hence the error is about ± 0.3 per cent.

30. **Simplification Method.** This method is widely used in practical calculations especially when we want to get a crude estimation of the result. The method includes such techniques as simplification of the original equation by dropping terms that are comparatively small, replacing slowly varying coefficients by constants and the

like. After this procedure is carried out we can arrive at an equation of one of the integrable types. Then integrating the equation we obtain a function which can be regarded as an approximate solution of the original equation. At any rate, such an approximation often correctly describes the qualitative character of the behaviour of the exact solution. After this "zero approximation" has been found we can often use it for determining corrections which compensate the simplifications we have performed, and thus we can find a "first approximation" and so on.

In case an equation involves parameters (e.g. masses, linear sizes of the objects under consideration etc.) we should take into account that the terms which we regard as being small can be different for different values of the parameters and therefore, generally, the simplification of an equation must be performed in different ways for differ ent values of the parameters involved. Besides, it is some-times ne cessary to break the interval of variation of the independent variable into several parts and simplify the equation on each of the parts by means of specific techniques pertaining to that very interval.

It is especially useful to simplify an equation in the way described above if we use certain simplifications in the very process of deducing the equation or if the degree of accuracy with which the quantities in question are determined is not sufficiently high. For in-stance, we should by all means drop such terms entering into an equation that are less than the admissible error of determination of its other terms.

For example, let us consider the problem

$$y'' + \frac{1}{1+0.1x}\, y + 0.2y^3 = 0, \ y(0) = 1,$$

$$y'(0) = 0, \quad 0 \leqslant x \leqslant 2 \tag{187}$$

The coefficient in y changing slowly, we replace the coefficient by its mean value (see Sec. XIV.5) which is equal to

$$\frac{1}{2-0} \int_0^2 \frac{1}{1+0.1x}\, dx = \frac{1}{2}\, \frac{\ln(1+0.1x)}{0.1} \Big|_0^2 = \frac{\ln 1.2}{0.2} = 0.911$$

Besides, let us drop the third summand which is comparatively small.

Thus we get the equation $y'' + 0.911y = 0$ whose solution satisfying the initial conditions is

$$y = \cos 0.954x \tag{188}$$

The form of this approximate solution justifies the procedure of dropping the last term in the equation because the ratio of the third term to the second term is of the order of $0.2y^2 < 0.2$ and there-

fore the first term and the second term must approximately "cancel out". Let us now bring in a correction connected with the last summand. To do this we substitute approximate solution (188) into the summand and retain the averaged value of the coefficient in y:

$$y'' + 0.911y = -0.2 \cos^3 0.954x =$$
$$= -0.05 \cos 2.86x - 0.15 \cos 0.954x$$

[here we have used formula (VIII.14)]. Now according to the methods of Sec. 18 we obtain the solution

$$y = 0.993 \cos 0.954\,x - 0.079x \sin 0.954x + 0.007 \cos 2.86x$$

satisfying the initial condition (check up the calculations!).

The difference between the last result and zero approximation (188) is not large and therefore our conclusions concerning the roles of different summands entering into equation (187) are confirmed again. At the same time we see that the third term of equation (187) has introduced a certain correction into the solution. [Think how we can determine the correction that takes into account the variability of the coefficient in y entering into equation (187).]

Considerations of the above type are often not sufficiently rigorous and may lead to mistakes. Therefore they should be applied in accordance with common sense. Then, comparatively often, they nevertheless result in approximate solutions that can be used for practical purposes.

31. Euler's Method. Now we proceed to study some methods of numerical integration of differential equations. These methods are particularly applicable to the cases when none of the above methods of "approximate analytical integration", that is methods of constructing approximate formulas for solutions, turns out to be ineffective, especially if a solution must be calculated with a great accuracy for a large interval of variation of the argument. Besides, the methods are used when equations are solved by means of electronic computers.

It is often advisable to combine methods of approximate and numerical integration. For instance, if an initial condition for the equation

$$y'' + (1 + e^{-x})\,y = 0$$

is given we can apply Taylor's formula for small values of x (see Sec. 24), one of the methods of numerical integration for intermediate values of x and, finally, we can simply drop the term e^{-x} for large values of x.

We shall consider four methods of numerical solution of first-order differential equations which are most frequently used. The methods are easily extended to systems of first-order equations to which we can also reduce higher-order equations. In courses on

approximate calculations one can find some other methods. We especially recommend [3], [9], [11] and [34].

Euler's method is visual and simple but not sufficiently effective. The reader must understand it well because many important and effective methods used in different branches of mathematics are essentially the development of the Euler method.

Euler's method is based on the procedure of direct replacement of the derivative entering into a differential equation by the ratio of finite differences (difference quotient) which were considered in Sec. V.7. Suppose we have an initial-value problem of the form

$$y' = f(x, y), \quad y(x_0) = y_0 \tag{189}$$

For simplicity's sake, let us take a constant step h along the x-axis. We introduce the notation

$$x_0 + h = x_1, \quad x_0 + 2h = x_2, \quad x_0 + 3h = x_3, \ldots$$

Approximate values $y(x_k)$ will be designated as y_k. To find the values we replace the derivative in the equation by the difference quotient. Hence we have

$$\frac{\Delta y_k}{\Delta x} = f(x_k, y_k)$$

i.e.

$$\frac{y_{k+1} - y_k}{h} = f(x_k, y_k)$$

and hence

$$y_{k+1} = y_k + f(x_k, y_k) h \tag{190}$$

Beginning with y_0 and making k assume, in succession, the values $k = 0, 1, 2, \ldots$, we apply formula (190) and compute the values

$$y_1 = y_0 + f(x_0, y_0) h, \quad y_2 = y_1 + f(x_1, y_1) h, \ldots$$

Euler's method has a simple geometric meaning which is illustrated in Fig. 301 where the integral curves are also depicted. We see that the geometric significance of the method lies in the fact that we draw the line segment $M_0 M_1$ tangent to the desired integral curve through the point M_0 instead of the integral curve itself which is unknown. In doing so we follow the direction of the direction field at the point M_0. We likewise draw the corresponding line segment through the point M_1 according to the direction of the tangent prescribed by the direction field at M_1 and so on. Thus we obtain Euler's broken line (polygonal line) which approximately represents the integral curve that would appear if the step h were infinitesimal, that is if we continuously "corrected" the direction of the broken line.

The order of the error of Euler's method can be easily estimated. Indeed, taking advantage of formula (190), we can replace the increment of the solution by its differential $y'_h \Delta x = f(x_h, y_h) h$. This leads to an error of the order of h^2 [see formula (IV.49)]. When we construct the solution on an interval (x_0, x) and break the interval into n parts we have $h = \frac{x - x_0}{n}$, and therefore the resultant error will be of the order of $nh^2 = \frac{(x - x_0)^2}{n}$. Hence, to decrease the error 10 times, that is to determine one more decimal digit, we must
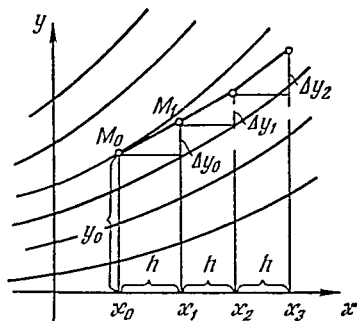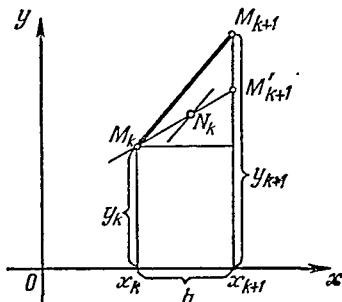
Fig. 301   Fig. 302

also increase the number of points of division 10 times which leads to a considerable increase of the amount of calculations. Here lies the disadvantage of the method.

There is a specific feature of Euler's method which is also characteristic of other methods of numerical integration of differential equations. We have already noted (see Sec. 7) that a solution of a differential equation may approach infinity for a finite value of x when it is continued along the x-axis. But at the same time it is clear that an approximate solution constructed in accordance with Euler's method remains finite for all values of x. To describe the behaviour of the solution in such a case correctly we can apply the following technique: if we see that there is a considerable increase of the solution in its absolute value we can perform the substitution $y = \frac{1}{z}$ in the differential equation. Then if further integration shows that z passes through a zero value attained at a certain point $x = \alpha$, this means that $|y(\alpha)| = \infty$.

32. Runge-Kutta Method. We now demonstrate a simpler variant of this method which specifies Euler's method. Suppose that an approximate value y_h of a solution for $x = x_h$ has already been computed. Then we can find y_{h+1} by calculating in accordance with

the formula

$$f_h = f(x_h, y_h), \quad \alpha_h = f\left(x_h + \frac{h}{2}, \ y_h + \frac{f_h h}{2}\right),$$

$$y_{h+1} = y_h + \alpha_h h \tag{191}$$

The geometric meaning of these calculations is illustrated in Fig. 302. Namely, every subsequent segment $M_h M_{h+1}$ of the broken line approximating the integral curve is constructed in the following way. We first draw the line segment $M_h M'_{h+1}$ according to the slope f_h of the direction field at the point M_h. just as in Euler's method. But here we do not limit ourselves to the result thus obtained and proceed to determine the slope α_h of the field at the midpoint N_h of the segment and to draw the new segment $M_h M_{h+1}$ in this direction. Thus, in this way we specify the slopes of the segments of the broken line approximating the integral curve.

Even the geometric illustration shows that this method is more accurate than Euler's method because here we take into account the change of the slope of the field along the interval $x_h \leqslant x \leqslant x_{h+1}$.

This can also be confirmed by calculations. By Taylor's formula (see Sec. XII.6) we get

$$\alpha_h = f(x_h, y_h) + f'_x(x_h, y_h)\frac{h}{2} + f'_y(x_h, y_h)\frac{f_h h}{2} + O(h^2)$$

where $O(h^2)$ designates a quantity which is bounded in comparison with h^2 (compare with Sec. III.11). As in Sec. 24, we find

$$y'_h = f(x_h, y_h) = f_h, \quad y''_h = f'_x(x_h, y_h) + f'_y(x_h, y_h) y'_h,$$

$$y_{h+1} = y_h + \alpha_h h = y_h + f(x_h, y_h) h + f_x(x_h, y_h)\frac{h^2}{2} +$$

$$+ f'_y(x_h, y_h) f_h \frac{h^2}{2} + O(h^3) = y_h + y'_h h + \frac{y''_h}{2} h^2 + O(h^3)$$

$$\tag{192}$$

But the exact value of the solution satisfying the condition $y(x_h) = y_h$ is equal to

$$y(x_h + h) = y_h + y'_h h + \frac{y''_h}{2} h^2 + O(h^3) \tag{193}$$

[see Taylor's formula (IV.50)]. Comparing formulas (192) and (193) we see that the values $y(x_h + h)$ and y_{h+1} can differ only in terms whose order of smallness is not less than that of h^3. From this, as in the end of Sec. 31, we conclude that the resultant error is of the order of $\frac{1}{n^2}$ or, which is the same, of the order of h^2. Hence, if we increase the number of points of division 10 times the degree of accuracy will increase 100 times.

A still more precise result will be obtained if we calculate according to the following scheme:

$$f_k = f(x_k,\ y_k), \quad \alpha_k = f\left(x_k + \frac{h}{2},\ y_k + \frac{f_k h}{2}\right),$$

$$\beta_k = f\left(x_k + \frac{h}{2},\ y_k + \frac{\alpha_k h}{2}\right), \quad \gamma_k = f(x_k + h,\ y_k + \beta_k h),$$

$$y_{k+1} = y_k + \frac{1}{6}(f_k + 2\alpha_k + 2\beta_k + \gamma_k)\, h$$

Calculations similar to (192) show that here the error made at every step does not exceed a quantity of the order of h^5 and therefore the resultant error is of the order of h^4. Consequently, if we increase the number of points of division 10 times the degree of accuracy will increase 10,000 times.

33. Adams Method. This method was introduced in 1883 by the English astronomer J. C. Adams (1819-1892). The method is based on Newton's second interpolation formula (V.29) which is applied to the derivative $y'(x)$ of the solution beginning with a certain value $x_h = x_0 + kh$ of the argument:

$$y'(x) = y'_k + \Delta y'_{k-1}\, \frac{x-x_k}{h} + \frac{\Delta^2 y'_{k-2}}{2!}\, \frac{x-x_k}{h}\left(\frac{x-x_k}{h} + 1\right) +$$

$$+ \frac{\Delta^3 y'_{k-3}}{3!}\, \frac{x-x_k}{h}\left(\frac{x-x_k}{h} + 1\right)\left(\frac{x-x_k}{h} + 2\right) \qquad (194)$$

In applying the formula we have substituted $x_k - x = -(x - x_k)$ for $t = x_{k+1} - x$ and, respectively, y'_k for the value y_{k+1}. Besides, we have replaced the sign of approximate equality by the sign of exact equality although formula (194) is, of course, approximate, and its error is of the order of $\Delta^4 y'$, that is of the order of h^4 (see Sec. V.7). Integrating formula (194) from x_k to $x_{k+1} = x_k + h$ (with the help of the substitution $\frac{x-x_k}{h} = s$) we receive

$$y_{k+1} = y_k + \left(y'_k + \frac{1}{2}\Delta y'_{k-1} + \frac{5}{12}\Delta^2 y'_{k-2} + \frac{3}{8}\Delta^3 y'_{k-3}\right) h \qquad (195)$$

(check up the calculations!).

The error of formula (195) is the result of the integration of the error of formula (194) and therefore it is of the order of h^5 (why is it so?).

Formula (195) is utilized in the following way. First we find the values

$$y_1 = y(x_0 + h), \quad y_2 = y(x_0 + 2h) \quad \text{and} \quad y_3 = y(x_0 + 3h)$$

by means of some other method, for instance, by Taylor's formula (see Sec. 24) or by the Runge-Kutta method (see Sec. 32). Then we

compute the corresponding values

$$y_0' = f(x_0, y_0), \quad y_1' = f(x_1, y_1), \quad y_2' = f(x_2, y_2)$$
$$\text{and} \quad y_3' = f(x_3, y_3)$$

This enables us to determine

$$\Delta y_0' = y_1' - y_0', \ \Delta y_1', \ \Delta y_2', \quad \Delta^2 y_0' = \Delta y_1' - \Delta y_0', \ \Delta^2 y_1',$$
$$\Delta^3 y_0' = \Delta^2 y_1' - \Delta^2 y_0'$$

Further, putting $k = 3$ in formula (195) we calculate y_4, and then using this value we find $y_4' = f(x_4, y_4)$, $\Delta y_3' = y_4' - y_3'$, $\Delta^2 y_2'$ and $\Delta^3 y_1'$. Then putting $k = 4$ in formula (195) we calculate y_5. After that we use y_5 for finding $y_5' = f(x_5, y_5)$ etc. The calculations are performed according to the following scheme:

34. Milne's Method. This method can be obtained by means of Newton's first interpolation formula (V.27). It is one of the most effective methods. Here we give only the final result, that is the scheme for performing calculations implied by the method (which was introduced in 1926).

The calculations are performed according to the following formulas:

$$\left. \begin{aligned} \overline{y}_{k+1} &= y_{k-3} + \frac{4h}{3}\left(2y_{k-2}' - y_{k-1}' + 2y_k'\right) \\ \overline{y}_{k+1}' &= f(x_{k+1}, \overline{y}_{k+1}) \\ y_{k+1} &= y_{k-1} + \frac{h}{3}\left(y_{k-1}' + 4y_k' + \overline{y}_{k+1}'\right) \end{aligned} \right\} \quad (k = 3, 4, 5, \ldots) \quad (196)$$

where $y_i' = f(x_i, y_i)$. Here, as in the Adams method, the values y_0, y_1, y_2, y_3 should be found in advance by means of some other method. After the values have been found we put $k = 3$ in formula (196) and calculate $\overline{y}_4, \overline{y}_4', y_4$, in succession. Then, putting $k = 4$

we find $\overline{y}_5$, $\overline{y}_5'$, y_5 and so on. The values y_4, y_5, y_6, ... thus determined are the approximate values of the solution $y(x)$ for $x = x_4$, x_5, x_6, ... where $x_i = x_0 + ih$.

It turns out that the absolute error which occurs when we calculate y_{k-1} according to this method is approximately equal to

$$\frac{|y_{k+1} - \overline{y}_{k+1}|}{29}$$

Therefore when performing the calculations we can simultaneously check whether the error lies within the limits of the degree of accuracy we have chosen for our calculations. If we see that at a certain stage of our calculations the error falls outside the prescribed limits we must decrease (from the corresponding value of x onwards) the step h taking into account that the resultant error of the method is of the order of h^4.

Multiple Integrals

§ 1. Definition and Basic Properties of Multiple Integrals

1. Some Examples Leading to the Notion of a Multiple Integral.
We now consider a solid (Ω) with the density of mass distribution ρ. The density ρ can be variable, that is different at different points of the solid. Let the function $\rho = \rho\,(M)$ (where M is a point of Ω) be known and let it be necessary to determine the whole mass m of the solid. An analogous problem for the case of a linear mass distribution was solved in Secs. XIV.1, 2. Let the reader read these sections again before proceeding to study multiple integrals.

The spatial case is treated quite similarly. Let us mentally divide the region (Ω) into n parts (subregions) ($\Delta\Omega_1$), ($\Delta\Omega_2$), ..., ($\Delta\Omega_n$), as in Fig. 303. Let the symbols ($\Delta\Omega_k$) ($k = 1, \ldots, n$) designate the parts themselves, and let $\Delta\Omega_k$ designate

Fig. 303

their volumes. Now, choose an arbitrary point M_k ($k = 1, \ldots, n$) in each of the subregions ($\Delta\Omega_k$). The points $M_1, M_2, \ldots, M_n$ are also shown in Fig. 303 which represents a spatial picture. If the parts ($\Delta\Omega_k$) are sufficiently small we can regard the density as being constant within each of the parts without an essential error. Then the mass $m_{(\Delta\Omega_1)}$ of the first part ($\Delta\Omega_1$) can be computed as the product of the density by the volume, i.e. as $\rho\,(M_1)\,\Delta\Omega_1$. The mass $m_{(\Delta\Omega_2)}$ of the second part is found similarly and so on. Thus, we obtain
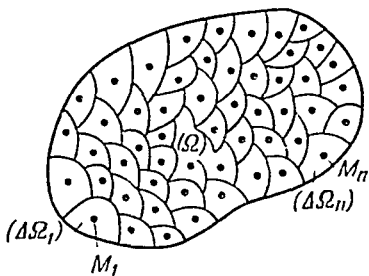
$$m_{(\Omega)} \approx \rho\,(M_1)\,\Delta\Omega_1 + \rho\,(M_2)\,\Delta\Omega_2 + \ldots + \rho\,(M_n)\,\Delta\Omega_n =$$
$$= \sum_{k=1}^{n} \rho\,(M_k)\,\Delta\Omega_k$$

This is an approximate equality since the densities of the parts are nevertheless variable. But the smaller the parts, the greater the accuracy. Hence, passing to the limit, as $\Delta\Omega_k \to 0$ ($k = 1, \ldots \ldots, n$), we obtain the exact equality

$$m_{(\Omega)} = \lim \sum_{k=1}^{n} \rho(M_k) \Delta\Omega_k \qquad (1)$$

The limit is taken here in a process in which not only the volumes but also all the linear sizes of the parts of the partitions tend to zero. Besides, it is supposed that the limit does not depend on the way of partitioning (Ω) into subregions.

Reasoning in a similar way we can conclude that if an electric charge is distributed over a solid (Ω) with density σ the magnitude q of the charge is found by means of the formula

$$q = \lim \sum_{k=1}^{n} \sigma(M_k) \Delta\Omega_k \qquad (2)$$

where the notation is understood as before.

A mass or a charge can be distributed not only in a volume but also over a surface or a curve. When we say that a mass or a charge is distributed over a surface this, of course, means that one of the dimensions of the domain of space in which the quantity is distributed is very small relative to the other two dimensions. The distribution along a curve is understood similarly. Formulas (1) and (2) remain true in the latter cases if the density ρ (or σ) is understood as a surface (areal) density (i.e. mass or charge per unit area) or as a linear density (i.e. related to unit length) and $\Delta\Omega_k$ designates the area or the length of the part $(\Delta\Omega_k)$ ($k = 1, \ldots, n$), respectively. In the general case we call $\Delta\Omega_k$ the *measure of the subregion* $(\Delta\Omega_k)$ understanding it as volume, area or length depending on whether we consider spatial regions, surfaces or curves.

2. **Definition of a Multiple Integral.** The similarity between formulas (1) and (2) indicates the advisability of the general definition of a multiple integral given below. For definiteness. let us consider integrals over three-dimensional regions. The measure of such a region is understood as its volume.

Suppose we are given a bounded (finite) region (Ω) in space. Let a function $u = f(M)$ be defined over (Ω) and let the value $f(M)$ of the function be finite at each point M of the region. To compose an integral sum we arbitrarily break up the region (Ω) into subregions $(\Delta\Omega_1)$, $(\Delta\Omega_2)$, $\ldots$, $(\Delta\Omega_n)$ and take an arbitrary point M_k ($k = 1, \ldots, n$) in each of them. Then taking the values $f(M_1)$, $f(M_2)$, $\ldots$, $f(M_n)$ of the function f assumed at the points M_1,

$M_2, \ldots, M_n$ we write down the integral sum

$$\sum_{k=1}^{n} u_k \, \Delta\Omega_k = \sum_{k=1}^{n} f(M_k) \, \Delta\Omega_k \tag{3}$$

where $\Delta\Omega_k$ ($k = 1, \ldots, n$) designates, as before, the volume of the subregion $(\Delta\Omega_k)$.

The limit of the integral sum taken in a process in which all the linear sizes of the subregions entering into the partitions of the region (Ω) are unlimitedly decreased is called the integral of the function f over the region (Ω). Denoting the integral by the symbol $\int_{(\Omega)} u \, d\Omega$ we can write

$$\int_{(\Omega)} u \, d\Omega = \int_{(\Omega)} f(M) \, d\Omega = \lim \sum_{k=1}^{n} f(M_k) \, \Delta\Omega_k \tag{4}$$

(compare this with the basic definitions given is Secs. XIV.2 and XIV.22). (Ω) is called the region (domain) of integration.

Consequently, formulas (1) and (2) can be rewritten as

$$m = \int_{(\Omega)} \rho \, d\Omega \quad \text{and} \quad q = \int_{(\Omega)} \sigma \, d\Omega$$

As in Sec. XIV.2, we can integrate both continuous and discontinuous functions. The existence of limit (4) can be proved for any finite function (under some additional conditions) defined in a finite region without referring to the physical meaning of an integral (by means of purely mathematical considerations). Besides, it is not the boundedness of the region that is essential for such a proof but the boundedness of its measure. Fig. 264 represents an example of a region which extends to infinity ·but has a finite measure.

The definition of an integral taken over a surface (which can be plane or curvilinear) or along a curve is formulated quite similarly. In these definitions we must, of course, take the areas or the lengths of the subregions instead of volumes when forming a partition of the region. In particular, an integral of this type along a curve is nothing but a line integral of the first type taken with respect to arc length which was studied in Sec. XIV.22. Integrals over a volume and over a surface are referred to as **multiple integrals**. The former are also called **triple integrals** and the latter are called **double integrals**. These terms will be explained in § 3.

3. Basic Properties of Multiple Integrals. The basic properties of a definite integral proved in Secs. XIV.4, 5 are implied by the definition of an integral as the limit of the integral sum (see Sec. XIV.2). Therefore we can easily extend these properties to multiple integrals. We enumerate them here.

1. The integral of a sum equals the sum of the integrals of the summands (the same is true for the difference):

$$\int\limits_{(\Omega)} (u_1 \pm u_2)\, d\Omega = \int\limits_{(\Omega)} u_1\, d\Omega \pm \int\limits_{(\Omega)} u_2\, d\Omega$$

2. A constant factor can be taken outside the sign of integration:

$$\int\limits_{(\Omega)} Cu\, d\Omega = C \int\limits_{(\Omega)} u\, d\Omega \quad (C = \text{const})$$

3. The theorem on a partition of the region of integration: for any partition of the region (Ω) into parts the integral over the whole region is equal to the sum of the integrals over the parts. For definiteness, if (Ω) is divided into the parts (Ω_1) and (Ω_2) we have

$$\int\limits_{(\Omega)} u\, d\Omega = \int\limits_{(\Omega_1)} u\, d\Omega + \int\limits_{(\Omega_2)} u\, d\Omega$$

4. The integral of unity is equal to the measure of the region of integration:

$$\int\limits_{(\Omega)} d\Omega = \Omega$$

5. If the region of integration degenerates, that is its measure turns into zero. the integral itself becomes equal to zero.

When formulating properties 4 and 5 we speak about the measure of a region (see Sec. 1) understanding it as volume, area or length depending on whether we consider triple, double or line integrals.

6. If the variables in question have certain dimensions then

$$\left[\int\limits_{(\Omega)} u\, d\Omega \right] = [u] \cdot [\Omega]$$

7. *The case of symmetry.* If the domain of integration can be divided into two symmetric parts and if the integrand takes equal values at the corresponding points belonging to these parts, the integral over the whole domain is equal to the doubled integral over each of these parts. If the integrand is multiplied by -1 when we pass from any point belonging to one of the parts to the symmetric point in the other part the integral over the whole region is equal to zero.

It is sometimes possible to break the domain of integration into a greater number of equal parts in order to reduce a given integral to an integral over a domain of a simpler form than the original domain.

8. It is allowable to integrate inequalities: if $u_1 \leqslant u_2$ then

$$\int\limits_{(\Omega)} u_1 \, d\Omega \leqslant \int\limits_{(\Omega)} u_2 \, d\Omega \tag{5}$$

The last inequality turns into the strict equality if and only if $u_1 \equiv u_2$ provided both functions u_1 and u_2 are continuous. But for the case of discontinuous functions integrals (5) can nevertheless coincide even when the identity $u_1 \equiv u_2$ is violated at points belonging to degenerated subregions which have a zero measure because such a violation does not affect the value of the integral (compare this with the corresponding property of an integral over a line segment).

9. An integral satisfies the inequalities

$$u_{min}\Omega \leqslant \int\limits_{(\Omega)} u \, d\Omega \leqslant u_{max}\Omega \tag{6}$$

10. Inequalities (6) are connected with the notion of the mean value $\overline{u}$ of a function u over a region (Ω) which is defined by means of the formula

$$\int\limits_{(\Omega)} \overline{u} \, d\Omega = \int\limits_{(\Omega)} u \, d\Omega \quad (\overline{u} = \text{const})$$

similar to that given in Sec. XIV.5. Thus, we have

$$\overline{u} = \frac{1}{\Omega} \int\limits_{(\Omega)} u \, d\Omega \quad \text{and} \quad \int\limits_{(\Omega)} u \, d\Omega = \overline{u}\Omega$$

Inequalities (6) imply that $u_{min} \leqslant \overline{u} \leqslant u_{max}$.

All these properties can be visually illustrated if we regard the function u as the density of a mass distribution and the integral as the mass itself.

11. There is an inequality of the form

$$\left| \int\limits_{(\Omega)} u \, d\Omega \right| \leqslant \int\limits_{(\Omega)} |u| \, d\Omega$$

which is similar to that given at the end of Sec. XIV.5.

4. Methods of Applying Multiple Integrals. There are two basic schemes of applying multiple integrals to physical problems (compare with Sec. XIV.6). The first one is based on representing the quantity in question in an approximate form of integral sum (3) in which we then pass to the limit, as it was shown in Sec. 1. The second scheme is based on composing the "element" (differential) of the quantity. We now briefly discuss the latter scheme (we shall dwell in more detail on the scheme in § 2).

Suppose we are interested in a quantity q which corresponds, for definiteness, to a spatial domain (Ω) [the situation is similar to that considered in Sec. 1 where we had a mass m or a charge q distributed over a three-dimensional region (Ω)]. Let us form the expression $dq = \varphi(M) \, d\Omega$ which approximately describes an infinitesimal portion of the quantity q corresponding to an infinitesimal volume $d\Omega$ placed at an arbitrary point M. This expression possesses the following properties: (1) it is directly proportional to the volume and (2) it differs from the true value Δq of the quantity q corresponding to the portion $d\Omega$ in an infinitesimal term of higher order of smallness relative to Δq. Now, summing all the quantities dq over all the "elements of volume" $d\Omega$ within the domain (Ω) we obtain

Fig. 304

$$q = q_{(\Omega)} = \int_{(\Omega)} \varphi(M) \, d\Omega \qquad (7)$$

As an example of the first method, let us consider the expression of the static moment (moment of mass) of a material body with respect to a plane (P). As is known from mechanics, the static moment of a finite system of material points relative to a plane (P) is expressed by the formula

$$S_{(P)} = \sum_k m_k z_k$$

where m_k is the mass of the kth point $(k = 1, 2, \ldots, n)$ and z_k is the coordinate of m_k reckoned along an axis drawn perpendicularly to the plane (P) (see Fig. 304).

If the mass is distributed over a spatial region (Ω) we divide the region into n parts $(\Delta\Omega_1), (\Delta\Omega_2), \ldots, (\Delta\Omega_n)$ and consider an approximate model in which the mass of each of the parts is concentrated at one of its points. This yields the approximate expression

$$S_{(P)} \approx \sum_{k=1}^n \rho_k z_k \, \Delta\Omega_k$$

which can be written in more detail as

$$S_{(P)} \approx \sum_{k=1}^n \rho(M_k) \, z(M_k) \, \Delta\Omega_k$$

Passing to the limit we thus obtain

$$S_{(P)} = \int_{(\Omega)} \rho z \, d\Omega \qquad (8)$$

The above calculations correspond to the first method mentioned at the beginning of this section.
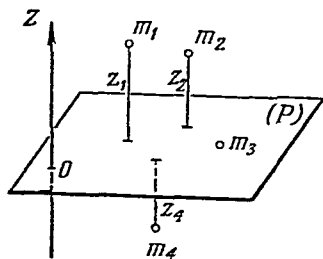
If we wanted to use the second method we should write the expression of the element of moment of mass:

$$dS_{(P)} = \rho z \, d\Omega$$

Summing, we should deduce the same formula (8).

Knowing the static moment we can readily find the coordinate z of the **centre of gravity** of the solid in question:

$$z_c = \frac{S_{(P)}}{m} = \frac{\int\limits_{(\Omega)} \rho z \, d\Omega}{\int\limits_{(\Omega)} \rho \, d\Omega}$$

The expression is simplified in the case when the solid is homogeneous, that is when $\rho = \text{const}$. Then the centre of gravity is referred to as the geometric centre of gravity, and we have

$$z_c = \frac{\rho \int\limits_{(\Omega)} z \, d\Omega}{\rho \int\limits_{(\Omega)} d\Omega} = \frac{1}{\Omega} \int\limits_{(\Omega)} z \, d\Omega \qquad (9)$$

The other coordinates of the centre of gravity of the body (Ω) are found similarly. The static moments and the coordinates of the centre of gravity of a plane geometric figure with respect to a straight line lying in the plane are found in like manner.

5. Geometric Meaning of an Integral Over a Plane Region. Such an integral, unlike other integrals defined in Sec. 2, can be directly interpreted geometrically. Its geometric meaning is similar to that of an ordinary definite integral considered in Sec. XIV.2. Let us be given an integral of the form

$$I = \int\limits_{(\Omega)} u \, d\Omega \qquad (10)$$

Fig. 305

where (Ω) is a domain lying in a plane (P) (see Fig. 305). Let us draw the u-axis perpendicularly to the plane and construct a line segment of length u (M) parallel to the u-axis and passing through a point M belonging to the domain (Ω). For simplicity's sake we now consider positive values of u; then the segment is drawn in the positive direction of the u-axis and the end-point N of the segment lies above the plane (P). If $u < 0$ the segment is drawn in the negative direction of the u-axis. When the point M runs throughout the domain (Ω) the corresponding point N describes a surface (S) which is the graph of the integrand (such a graph was constructed in Fig. 190). The surface (S) together with the plane figure (Ω) and
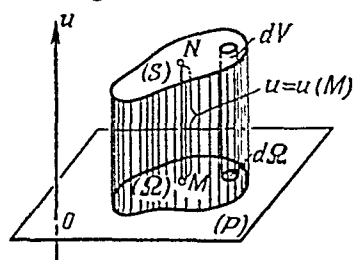
the cylindrical surface formed by the line segment parallel to the u-axis and drawn through each point of the contour bordering the domain (Ω) bound a cylindrical body.

The geometric meaning of integral (10) *lies in the fact that it is equal to the volume of the cylindrical body.* Indeed, the element of volume corresponding to a surface element $d\Omega$ of the domain (Ω) (see Fig. 305) containing a point M can be regarded as a right cylinder with base $d\Omega$ and height u (M) to within infinitesimals of higher order. Hence, this volume is approximately equal to $dV = u \, d\Omega$. Summing up these elements of volume we arrive at the formula

$$V = \int\limits_{(\Omega)} u \, d\Omega = I$$

which is what we set out to prove.

If u assumes negative values as well the volumes of the parts of the body lying under the plane (P) enter into the result with the sign minus (this situation is quite analogous to the one considered in Sec. XIV.2).

By analogy with formula (XIV.28), we can pass from the volumes of cylindrical bodies to the volume of a body having an arbitrary form. If (Ω) is the projection of the body on a plane (P) and $h = = h$ (M) is the length of the line segment formed by the intersection of the body with a straight line which is perpendicular to (P) and passes through the current point M of the domain (Ω), we have the expression

$$V = \int\limits_{(\Omega)} h \, d\Omega$$

for the volume of the body.

§ 2. Two Types of Physical Quantities

6. **Basic Example. Mass and Its Density.** We now consider a material body whose density can be variable in the general case. Let us disregard the molecular structure of the body and consider its mass to be continuously distributed in space. According to this model such a body has a certain density ρ at each of its points M, and hence ρ is a function of a point of the form $\rho = \rho$ (M) (see Sec. IX.9). In contrast to it, the mass cannot be regarded as a function of a point because the mass of a separate point is equal to zero. The mass is a quantity which is distributed in space which means that to each region (Ω) (regarded as being mentally taken out of the space) there corresponds a certain value $m_{(\Omega)}$ of the mass. As before, (Ω) is regarded here as a symbol designating the region itself, not its volume.

The connection between the mass and the density is as follows. Let the value $m_{(\Omega)}$ of the mass corresponding to each domain (Ω) in space be known. Then the ratio

$$\rho_{av} = \frac{m_{(\Omega)}}{\Omega}$$

is referred to as the mean (average) density in (Ω). The symbol Ω entering into the ratio designates, as in Sec. 1, the volume of the domain (Ω). Hence, Ω is a quantity having the dimension of volume. [By the way, we shall sometimes refer to (Ω) as a volume when there is no danger of misunderstanding.]

To obtain the density at a certain point M we must pass to the limit (compare with Sec. IV.1) by making (Ω) contract to the point M:

$$\rho(M) = \lim_{(\Omega) \to M} \rho_{av} = \lim_{(\Omega) \to M} \frac{m_{(\Omega)}}{\Omega} \tag{11}$$

This process is analogous to that of calculating a derivative. When we write $(\Omega) \to M$ we mean that (Ω) is unlimitedly contracted to the point M. If $(\Omega) \to M$ we have, naturally, $\Omega \to 0$ but the requirement that $(\Omega) \to M$ is stronger than the condition $\Omega \to 0$ (why?). Thus, the density of mass distribution at a point is the mass of an infinitesimal volume related to unit volume.

Conversely, if the value $\rho(M)$ of the density is known for each point M the mass $m_{(\Omega)}$ corresponding to any part (Ω) of space can be found on the basis of Secs. 1, 2 as the integral

$$m_{(\Omega)} = \int\limits_{(\Omega)} \rho \, d\Omega$$

If we now take into account the molecular structure of the substance, i.e. if we return from our idealized model to reality, we cannot even mentally contract the volume (Ω) to a point unlimitedly. In this case, instead of formula (11), we must write the expression

$$\rho(M) = \frac{m_{(\Delta\Omega)}}{\Delta\Omega}$$

where $(\Delta\Omega)$ is a volume containing the point M which is regarded as being practically infinitesimal (see Sec. III.1). Consequently, the density of a real body at a point is the mean density in a volume which is sufficiently small relative to the dimensions of the body and at the same time sufficiently large relative to the molecular sizes. Hence, in these considerations we pass from the real discrete structure of a material body to its continuous model in which the density is obtained in the process of averaging based on the calculation of the mean density corresponding to the volumes whose sizes were indicated above.

38—0141

Further, when considering a continuous medium in our course, we shall always take the continuous model abstracting from the molecular structure of substance.

7. Quantities Distributed in Space. There is a number of physical quantities which are analogous to mass in many respects. They possess the properties considered in the above example of mass distribution. Examples of such quantities are an electric charge in a dielectric, quantity of heat, energy of an electromagnetic field and the like. There is a feature which is common to all such quantities, namely, they are all distributed in space. In the general case we say that a quantity q is distributed in space if to each part (Ω) mentally isolated from the space there corresponds a certain value $q_{(\Omega)}$ of the quantity. There is only one general requirement which we introduce here: the quantity must be additive, i.e. for any partition of (Ω) into parts the value of the quantity corresponding to (Ω) must be equal to the sum of the values of the quantity corresponding to the parts. Hence, if, for definiteness, (Ω) is divided into two parts (Ω_1) and (Ω_2) we must have $q_{(\Omega)} = q_{(\Omega_1)} + q_{(\Omega_2)}$.

A quantity distributed in space possesses a certain density at each point. Thus we can speak about the density of an electric charge, of field energy and so on. In the general case the density φ is defined by analogy with formula (11):

$$\varphi(M) = \lim_{(\Omega) \to M} \frac{q_{(\Omega)}}{\Omega} \tag{12}$$

The ratio under the sign of limit in (12) is the mean (average) density of the quantity q in the volume (Ω). The density $\varphi = \varphi(M)$ is then a function of a point. The density of the quantity q at a point M equals the value of q (corresponding to an infinitesimal region "placed at the point M") related to unit volume.

Conversely, if the density $\varphi(M)$ of a quantity q is known the quantity q itself is found by the methods described in Secs. 1, 2:

$$q_{(\Omega)} = \lim \sum_{k=1}^{n} \varphi(M_k) \Delta\Omega_k = \int_{(\Omega)} \varphi(M) \, d\Omega \tag{13}$$

where the limit is taken in the process in which the linear sizes of the parts forming partitions of the region are decreased unlimitedly.

In the general case q and φ can take on the values of any sign.

Let us rewrite formula (12) so as to stress that we deal with infinitesimal volumes. To do this we substitute $(\Delta\Omega)$ for Ω:

$$\frac{q_{(\Delta\Omega)}}{\Delta\Omega} \to \varphi(M), \quad \text{i.e.} \quad \frac{q_{(\Delta\Omega)}}{\Delta\Omega} = \varphi(M) + \alpha$$

where α is infinitesimal when $(\Delta\Omega) \to M$. This implies

$$q_{(\Delta\Omega)} = \varphi(M)\Delta\Omega + \alpha\Delta\Omega$$

Thus, the value of q corresponding to a small volume $(\Delta\Omega)$ is divided into two parts the first of which is directly proportional to the volume $\Delta\Omega$ while the other is of higher order of smallness. The former summand is therefore called the **differential** or the **element of the quantity** q (compare with Secs. IV.7, 8):

$$dq = \varphi(M)\Delta\Omega \tag{14}$$

This implies the physical meaning of dq: it is the value of q which would correspond to the volume $(\Delta\Omega)$ if the density were constant throughout $(\Delta\Omega)$ and were equal to the density at the point M. In reality, Δq, i.e. $q_{(\Delta\Omega)}$, does not equal dq in the general case and differs from it by an infinitesimal of higher order of smallness. Hence, Δq and dq are equivalent infinitesimals when $(\Delta\Omega) \to M$ (see Sec. III.8). If it is possible to neglect such infinitesimals of higher order we simply say that dq is the value of q corresponding to an infinitesimal volume $(\Delta\Omega)$. We also call it an infinitesimal mass, an infinitesimal charge etc. in such cases.

The volume can be regarded as a special case of a quantity q distributed in space, that is we can put $q_{(\Omega)} = \Omega$. The corresponding density is then equal to unity (as "a volume related to unit volume"). Hence, formula (14) implies

$$d\Omega = \Delta\Omega$$

in this case. Therefore formula (14) can be put down in the form

$$dq = \varphi(M)\, d\Omega \tag{15}$$

which is preferable (this resembles the corresponding formula in Sec. IV.9).

Thus, summing up, we can write the basic formulas connecting a quantity $q = q_{(\Omega)}$ distributed in space and the corresponding function of a point (density) $\varphi = \varphi(M)$ in the form

$$\varphi(M) = \frac{dq}{d\Omega}\Big|_M \quad \text{and} \quad q_{(\Omega)} = \int\limits_{(\Omega)} \varphi(M)\, d\Omega$$

These formulas enable us to pass from one representation of a quantity to the other in all cases. The density is found by means of differentiating the quantity and the quantity itself is found by integrating its density.

It should be noted that there are some quantities differing in their nature from such quantities as masses and charges which can also be considered to be distributed in space or over a surface or

a curve. For instance, the static moment or the moment of inertia of a material body satisfies the conditions formulated in the definition given at the beginning of Sec. 7 and thus they can be regarded as being distributed in space although they depend on the choice of a plane or of an axis. In practical problems we do not distinguish between such quantities belonging to different classes and simply write the expression dq according to the considerations given in Sec. 4 and then perform the summation (or integration) of the elements on the basis of the additivity law and thus arrive at an expression of form (7).

A quantity can be distributed not only over a volume but also over a surface (plane or curvilinear) or along a curve. All the results obtained in this section remain valid for these cases if we interpret (Ω) not as a three-dimensional domain mentally taken out of space but as a part of a surface (i.e. a region on a surface) or a part of a curve and understand Ω as the area or length of the part, that is as its measure (see Sec. 1).

§ 3. Computing Multiple Integrals in Cartesian Coordinates

8. Integral Over Rectangle. We now consider an integral

$$I = \int\limits_{(\Omega)} u \, d\Omega \tag{16}$$

where (Ω) is a rectangle bounded by coordinate lines of a Cartesian coordinate system arbitrarily chosen in a plane (see Fig. 306). The rectangle is described by inequalities $a \leqslant x \leqslant b$ and $c \leqslant y \leqslant d$
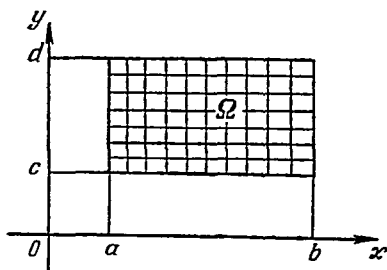
where a, b, c, d are some constants. When forming an integral sum S for integral (16) in the Cartesian coordinate system it is natural to break up (Ω) into parts by means of straight lines parallel to the coordinate axes which divide the interval $a \leqslant x \leqslant b$ into parts Δx_i and the interval $c \leqslant y \leqslant d$ into parts Δy_l. Let us denote by u_{ik} the value of the integrand $u = u(x, y)$ at a point belonging to the subregion adjoining the intersection of the ith

Fig. 306

vertical line with the kth horizontal line (see Fig. 306). (Let the reader pay attention to the fact that the numeration of u_{ik} does not coincide with the one used in the theory of matrices in Sec. XI where a_{ik} designates the element of a matrix lying at the intersection of the ith row of the matrix with its kth column.)

We then approximately have

$$I \approx S = \sum_{i,\,k} u_{ik}\Delta x_i \Delta y_k \qquad (17)$$

where the summation is extended over all the subregions (small rectangles). i.e. over all the values of i and k (for instance, $i = 1, 2, \ldots \ldots, m$ and $k = 1, 2, \ldots, n$).

A sum of form (17) with two summation indices is a two-dimensional integral sum. To compute it we must first perform summation with respect to k for a fixed i, that is to sum up the summands forming a column (the ith column) of the table

$$
\begin{array}{ccccc}
u_{11} & u_{21} & u_{31} & \cdots & u_{m1} \\
u_{12} & u_{22} & u_{32} & \cdots & u_{m2} \\
\cdot\;\cdot\;\cdot & \cdot\;\cdot\;\cdot & \cdot\;\cdot\;\cdot & \cdots & \cdot \\
u_{1n} & u_{2n} & u_{3n} & \cdots & u_{mn}
\end{array}
$$

for each fixed value of i ($i = 1, 2, \ldots, m$) and then perform the summation with respect to i. This results in

$$S = \sum_{i=1}^{m} \left(\sum_{k=1}^{n} u_{ik}\Delta x_i \Delta y_k \right) = \sum_{i=1}^{m} \left(\sum_{k=1}^{n} u_{ik}\Delta y_k \right) \Delta x_i \qquad (18)$$

where we have taken outside the brackets the common factor entering into the summands of the inner sum. The transition from a two-dimensional sum to a two-fold iterated sum, that is from (17) to (18), can also be performed in the reverse order: the first, inner, sum can be taken with respect to i and the outer, repeated, sum with respect to k.

If the divisions along the y-axis are sufficiently small the sum inside the brackets in (18) is close to the corresponding integral:

$$\sum_{k=1}^{n} u_{ik}\Delta y_k \approx \left(\int_{c}^{d} u\, dy \right)_i \qquad (19)$$

where the subscript i indicates that the value of x, which is fixed, is taken for the ith column of the above table. It follows that

$$S \approx \sum_{i=1}^{m} \left(\int_{c}^{d} u\, dy \right)_i \Delta x_i$$

But this is also an integral sum for function (19) which depends on x. Hence, if the divisions along the x-axis are also sufficiently small we can write

$$S \approx \int_{a}^{b} \left(\int_{c}^{d} u\,(x,\, y)\, dy \right) dx \qquad (20)$$

because the sum is close to integral (20).

In the process of decreasing the subregions of the partitions equalities (17) and (20) become more and still more accurate and turn into the precise relations in the limit. Consequently,

$$I = \int\limits_{(\Omega)} u \, d\Omega = \int\limits_a^b \left(\int\limits_c^d u(x, y) \, dy \right) dx \tag{21}$$

Thus, to compute an integral taken over a rectangle with sides parallel to the coordinate axes we can first perform the integration with respect to y, for a fixed x, as the variable of integration y varies within the rectangle (the **inner integration**) and then integrate the result of the first integration (which depends only on x) with respect to x within the corresponding limits of its variation (the **outer integration**).

The reverse order of passing from sum (17) to an iterated twofold sum (see above) would yield

$$\int\limits_{(\Omega)} u \, d\Omega = \int\limits_c^d \left(\int\limits_a^b u(x, y) \, dx \right) dy \tag{22}$$

Hence, when computing a double integral in Cartesian coordinates we have two ways of passing to a repeated (iterated) two-fold integral (as it will be shown in § 4, an analogous passage to an iterated integral can be performed in any coordinate system). It should be noted that one of the ways usually turns out to be more difficult for practical calculations whereas the other is simpler. The transition from one of these ways to the other is referred to as the **inversion of the order of integration**.

Fig. 307

Formula (21) can be readily interpreted geometrically. According to Sec. 5, integral (16) is equal to the volume of the solid depicted in Fig. 307. On the basis of Sec. XIV.10 we can compute the volume by integrating the cross section area shaded in Fig. 307. Hence, we obtain

$$\int\limits_{(\Omega)} u \, d\Omega = V = \int\limits_a^b S(x) \, dx = \int\limits_a^b \left(\int\limits_c^d u \, dy \right) dx$$

The geometric meaning of formula (22) is analogous to that of (21). Hence, formulas (21) and (22) can be proved in a simpler manner but we have given a more complicated method of proving since
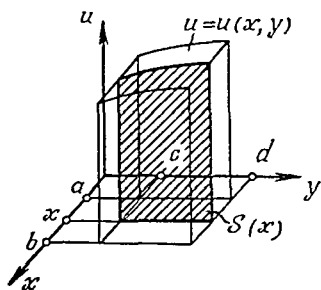
it can be automatically extended to multiple integrals of an arbitrary order.

Bearing in mind formulas (21) and (22) we sometimes denote the original integral (16) as

$$I = \iint\limits_{(\Omega)} u \, d\Omega \quad \text{or} \quad I = \iint\limits_{(\Omega)} u \, dx \, dy$$

meaning that $d\Omega = dx \, dy$ for a partition of the type shown in Fig. 306 when the divisions along the coordinate axes are sufficiently small.

The computation of a repeated integral with constant limits of integration of form (21) becomes particularly simple when the integrand is a product of two factors each of which depends only on one variable of integration. Namely, if $u \, (x, \, y) = f_1 \, (x) \, f_2 \, (y)$ we have

$$I = \int\limits_a^b \left(\int\limits_c^d f_1 \, (x) \, f_2 \, (y) \, dy \right) dx = \int\limits_a^b f_1 \, (x) \left(\int\limits_c^d f_2 \, (y) \, dy \right) dx =$$

$$= \int\limits_a^b f_1 \, (x) \, dx \cdot \int\limits_c^d f_2 \, (y) \, dy$$

Thus we have obtained the product of two one-dimensional integrals.

9. Integral Over an Arbitrary Plane Region. Let (Ω), entering into integral (16), be an arbitrary plane figure lying in the x, y-plane. For instance, take the domain depicted in Fig. 308. The considerations given in Sec. 8 can be transferred to this case with some slight changes. Namely, instead of integral (19) we arrive at an integral of the form

$$\int\limits_{y_1}^{y_2} u \, (x, \, y) \, dy = \int\limits_{\varphi_1(x)}^{\varphi_2(x)} u \, (x, \, y) \, dy$$

where $y = y_1 = \varphi_1 \, (x)$ and $y = y_2 = \varphi_2 \, (x)$ are, respectively, the equations of the upper and lower parts of the boundary of the domain (Ω). The contour bordering the figure (Ω) is divided into these two parts by the points A and B (see Fig. 308). Accordingly, the final result [which substitutes for formula (21) in this case] will be of the form

$$I = \int\limits_{(\Omega)} u \, d\Omega = \int\limits_a^b \left(\int\limits_{\varphi_1(x)}^{\varphi_2(x)} u \, (x, \, y) \, dy \right) dx \qquad (23)$$

Consequently, the limits of integration in the inner integral are variable in the general case; they depend on the variable of integration in the outer integral (i.e. on x in our case). The character of

this dependence is specified by the form of the contour. But the limits of integration in the outer integral are constant as before. They are specified by the maximal range of variation of x. Hence, we see that the rule given after formula (21) remains valid for a domain (Ω) of general form.

Here we can also invert the order of integration, that is perform the first integration with respect to x and the second integration
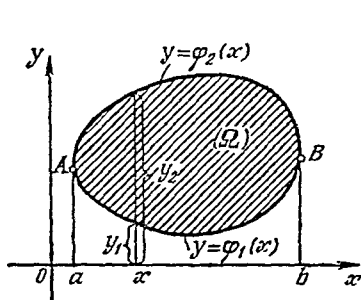
Fig. 308
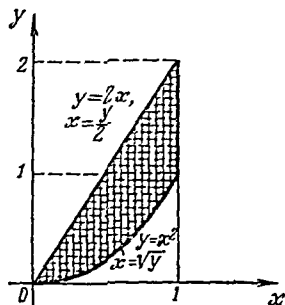
Fig. 309

with respect to y. Then in place of formula (22) we arrive at a formula of the form

$$\int\limits_{(\Omega)} u \, d\Omega = \int\limits_c^d \left(\int\limits_{\psi_1(y)}^{\psi_2(y)} u(x, y) \, dx \right) dy \tag{24}$$

[let the reader find out what c, d, $\psi_1(y)$, $\psi_2(y)$ are by examining Fig. 308].

It is sometimes necessary to break the domain of integration into several parts before setting up the limits of integration.

For example, let it be necessary to invert the order of integration in the integral

$$I = \int\limits_0^1 dx \int\limits_{x^2}^{2x} f(x, y) \, dy$$

which can be written at length as

$$I = \int\limits_0^1 \left(\int\limits_{x^2}^{2x} f(x, y) \, dy \right) dx \tag{25}$$

To do this we must first determine the geometric form of the domain of integration. In this case it is bounded by the lines $x = 0$, $x = 1$, $y = x^2$ and $y = 2x$ (see Fig. 309), and the first, inner, integration is performed along the line segments parallel to the y-axis, the segments being shown in continuous lines in Fig. 309.

After the inversion of the order of integration the inner integration will be carried out along the line segments parallel to the x-axis shown in dotted lines in Fig. 309. We see that after the order of integration has been inverted the inner integration is performed from the straight line $x = \frac{y}{2}$ to the parabola $x = \sqrt{y}$ for $y < 1$ and from the straight line $x = \frac{y}{2}$ to the straight line $x = 1$ for $y > 1$. The value of the variable y which corresponds to the division of the domain of integration into the two parts (in which the
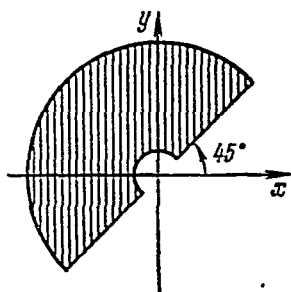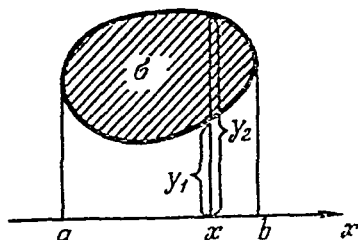
Fig. 310   Fig. 311

upper limits of the inner integrals differ) is found as the ordinate of the point of intersection of the parabola $y = x^2$ with the straight line $x = 1$, i.e. $y = 1$. Hence, after the inversion of the order of integration we obtain the sum of two integrals of the form

$$I = \int_0^1 dy \int_{\frac{y}{2}}^{\sqrt{y}} f(x, y)\, dx + \int_1^2 dy \int_{\frac{y}{2}}^1 f(x, y)\, dx$$

instead of formula (25).

In more complicated cases it is sometimes necessary to divide the domain of integration into a greater number of parts. For example, to set up the limits of integration in Cartesian coordinates for an integral taken over the domain shown in Fig. 310 it is necessary to break the domain into five parts (what are these parts?).

We now consider a simple example of an application of the double integral. By analogy with formula (9). we can easily deduce the formulas for the coordinates of the geometric centre of gravity of a plane figure (σ):

$$x_c = \frac{\iint\limits_{(\sigma)} x\, dx\, dy}{\sigma}, \qquad y_c = \frac{\iint\limits_{(\sigma)} y\, dx\, dy}{\sigma} \qquad (26)$$

where σ designates the area of the figure (σ). Let the whole figure (σ) entirely lie on one side of the x-axis (see Fig. 311). Then the second formula (26) can be rewritten in the form

$$\sigma \cdot y_c = \int\limits_a^b dx \int\limits_{y_1}^{y_2} y\, dy = \int\limits_a^b dx \left(\frac{y_2^2}{2} - \frac{y_1^2}{2} \right) =$$

$$= \frac{1}{2} \int\limits_a^b y_2^2\, dx - \frac{1}{2} \int\limits_a^b y_1^2\, dx$$

Multiplying both sides by 2π and recalling formula (XIV.35) of the volume of a solid of revolution we arrive at **Guldin's second theorem**: *if a homogeneous plane figure rotates about an axis lying*
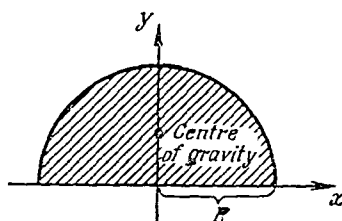
Fig. 312
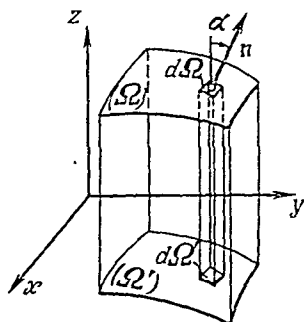
Fig. 313

in the plane of the figure and not intersecting it the volume of the solid of revolution thus obtained is equal to the product of the area of the figure by the distance covered by its centre of gravity. On the basis of the theorem we readily find the geometric centre of gravity of a semicircle of radius R (see Fig. 312):

$$\frac{4}{3} \pi R^3 = \frac{\pi R^2}{2} \cdot 2\pi y_c$$

that is

$$y_c = \frac{4}{3\pi} R = 0.425 R$$

10. Integral Over an Arbitrary Surface. Let us consider the integral

$$I = \int\limits_{(\Omega)} u\, d\Omega \tag{27}$$

taken over an arbitrary surface (Ω) which can be curvilinear in th general case (see Fig. 313). To compute it in Cartesian coordinat we must consider the projection of the surface (Ω) on one of th coordinate planes. For definiteness, let us take the projection (Ω) on the x, y-plane which we denote by (Ω').

Since the element (the area of an infinitesimal part) of a curvil near surface can be regarded as being plane to within infinitesima of higher order of smallness relative to the area, we have

$$d\Omega' = d\Omega \, |\cos \alpha \,| = d\Omega \, |\cos \widehat{(\mathbf{n}, z)} \,|$$

where $\mathbf{n}$ is a normal vector to the surface. It follows that

$$I = \int\limits_{(\Omega)} u \, d\Omega = \int\limits_{(\Omega')} u \, \frac{d\Omega'}{|\cos \widehat{(\mathbf{n}, z)}|} \tag{2$}$$

The last integral taken over the plane figure (Ω') is computed h means of the methods of Sec. 9.

Let the surface in question be represented by an equation of th form $z = f(x, y)$. Then, according to Sec. XII.2, the vector

$$\mathbf{n} = -\frac{\partial f}{\partial x} \, \mathbf{i} - \frac{\partial f}{\partial y} \, \mathbf{j} + \mathbf{k}$$

is directed along the normal to the surface at every point x, y, belonging to the surface. Hence (see Sec. VII.10) we have

$$\cos \widehat{(\mathbf{n}, z)} = \frac{\mathbf{n} \cdot \mathbf{k}}{|\mathbf{n}| \cdot |\mathbf{k}|} = \frac{1}{\sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + 1}}$$

Therefore, if integral (27) is given in the form

$$I = \int\limits_{(\Omega)} u(x, y, z) \, d\Omega$$

we obtain, on the basis of formula (28), the expression

$$I = \iint\limits_{(\Omega')} u(x, y, f(x, y)) \sqrt{1 + (f'_x)^2 + (f'_y)^2} \, dx \, dy$$

In particular, taking into account property 4 in Sec. 3 we deriv the formula for the area Ω of an arbitrary surface (Ω):

$$\Omega = \iint\limits_{(\Omega)} d\Omega = \iint\limits_{(\Omega')} \sqrt{1 + (z'_x)^2 + (z'_y)^2} \, dx \, dy$$

Here, as above, (Ω') is the projection of the surface (Ω) on the x, plane and $z = z(x, y)$ is the equation of the surface.

When projecting a surface, we sometimes have to divide it int several parts. The projection on the planes y, z or x, z and the co

responding computation of a surface integral are performed in a similar way when it is expedient. [Let the reader deduce the formula of $|\cos (\widehat{n, z})|$ for a surface represented by an equation of the form $F(x, y, z) = 0.$]

11. Integral Over a Three-Dimensional Region. Let us now consider an integral

$$I = \int\limits_{(\Omega)} u \, d\Omega$$

where (Ω) is a solid, that is a domain in space. We compute it following the procedure which was developed in Secs. 8 and 9 for an integral over a plane figure. The corresponding integral sum is now represented as a three-fold iterated sum. In the simplest case when (Ω) is a rectangular parallelepiped defined by the inequalities $a \leqslant x \leqslant b$, $c \leqslant y \leqslant d$ and $e \leqslant z \leqslant f$ we obtain, after passing to the limit in the integral sum, the formula

$$I = \int\limits_a^b dx \int\limits_c^d dy \int\limits_e^f u(x, y, z) \, dz$$

that is

$$I = \int\limits_a^b \Big(\int\limits_c^d \Big(\int\limits_e^f u(x, y, z) \, dz \Big) \, dy \Big) \, dx$$

By the way, it is possible to perform here the integration by inverting the order of integration in five different ways because there are six different combinations (permutations) of the differentials dx, dy, dz.

In the case of a domain of integration of a more general form the determination of the limits of integration will be more complicated. Suppose that we want to set up the limits of integration when integrating in the following order:

$$I = \int\limits_{(\Omega)} u \, d\Omega = \int dx \int dy \int u(x, y, z) \, dz \qquad (29)$$

Let the domain of integration be of the form shown in Fig. 314. Here the first (inner) integration is performed with respect to z within the domain (Ω) for fixed x and y. Therefore the limits of this integration are z_1 and z_2 (see Fig. 314), i.e. $\varphi_1(x, y)$ and $\varphi_2(x, y)$ where $z = \varphi_1(x, y)$ and $z = \varphi_2(x, y)$ are the equations of the upper and lower parts of the surface bordering the solid (Ω).

After the integration with respect to z and the substitution of the limits of integration have been performed the result of the first integration depends only on x and y. Now we pass to the projection

(Ω') of the solid (Ω) on the x, y-plane and perform the integration with respect to y (the second integration). When integrating with respect to y within the projection (Ω') we keep x fixed. Thus the limits of the second integration are $y_1 = \psi_1(x)$ and $y_2 = \psi_2(x)$, as it was described in Sec. 9. The result of the second integration will depend only on x. It should be integrated with respect to x over the maximal range of x, that is from a to b. This is the third, outer integration. Thus, after setting up the limits of integration we can put down integral (29) in the form

Fig. 314

$$\int\limits_{(\Omega)} u \, d\Omega =$$

$$= \int\limits_a^b dx \int\limits_{\psi_1(x)}^{\psi_2(x)} dy \int\limits_{\varphi_1(x, y)}^{\varphi_2(x, y)} u(x, y, z) \, dz$$

The reader should pay attention to the fact that the limits of integration in each integral depend only on those variables with respect to which the integration has not yet been performed. In particular, the limits of the outer integration cannot depend on the variables of integration and are constant.

We can similarly set up the limits of integration in the other five possible cases of inverting the order of integration. As in Sec. 9, we sometimes have to divide the domain of integration into several parts when setting up the limits of integration if the domain (Ω) is of a more complicated form.
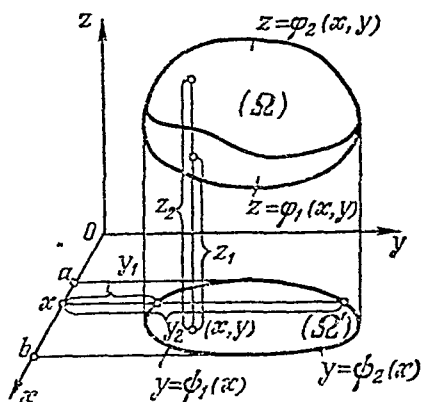
§ 4. Change of Variables in Multiple Integrals

12. Passing to Polar Coordinates in Plane. As in the case of a one-dimensional integral, we can introduce different variables of integration when computing a double integral. Here we shall consider a typical example of computing a double integral in polar coordinates. Let us take an integral of the form

$$I = \int\limits_{(\Omega)} u \, d\Omega$$

where (Ω) is a region in the x, y-plane which is depicted in Fig. 315. If it is necessary to perform the integration in polar coordinates we must divide the domain into parts by means of the coordinate

curves of the polar coordinate system, i.e. by the lines $\rho = \text{const}$ and $\varphi = \text{const}$ (see Sec. II.5), as it is shown in Fig. 315. Each of the elementary areas thus obtained can be regarded as being equal to a rectangle with sides $d\rho$ and $\rho\, d\varphi$ to within infinitesimals of higher order (why is it so?). Hence, we have

$$d\Omega = \rho\, d\rho\, d\varphi$$

Performing the summation over all the elementary areas we obtain

$$I = \iint\limits_{(\Omega)} u\rho\, d\rho\, d\varphi$$

where the integrand must be, of course, expressed as a function of ρ and φ. By analogy with Sec. 9, we set up the limits of integration and thus receive
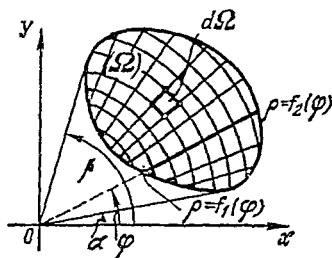
Fig. 315

$$\int\limits_{(\Omega)} u\, d\Omega = \int\limits_{\alpha}^{\beta} d\varphi \int\limits_{f_1(\varphi)}^{f_2(\varphi)} u\rho\, d\rho \quad (30)$$

The geometric meaning of the limits of integration is illustrated in Fig. 315.

Polar coordinates are particularly convenient for regions whose boundary consists of coordinate curves of the polar coordinate system because in such cases, when setting up the limits of integration, we obtain constant limits not only in the outer integral but also in the inner one. For example, after the limits of integration are set up, an integral taken over the domain shown in Fig. 310 will have the form

$$\int\limits_{\frac{\pi}{4}}^{\frac{5\pi}{4}} d\varphi \int\limits_{r_1}^{r_2} u\rho\, d\rho$$

13. Passing to Cylindrical and Spherical Coordinates. Let us take an integral

$$I = \int\limits_{(\Omega)} u\, d\Omega \quad (31)$$

where (Ω) is a domain of space. If it is necessary to perform the integration in cylindrical coordinates (see Sec. X.1) we have to divide the domain into parts by means of the coordinate surfaces of the cylindrical coordinate system, i.e. the surfaces $\rho = \text{const}$, $\varphi = \text{const}$

and $z =$ const. Then each of the elements of volume (see Fig. 316) can be regarded as being equal to the volume of the rectangular parallelepiped with dimensions $d\rho$, $\rho\,d\varphi$ and dz to within infinitesimals of higher order of smallness (relative to the element of volume). We suggest that the reader should verify this assertion. Consequently, we have

$$d\Omega = d\rho\cdot\rho\,d\varphi\cdot dz = \rho\,d\rho\,d\varphi\,dz$$

Therefore integral (31) takes the form

$$I = \iiint\limits_{(\Omega)} u\rho\,d\rho\,d\varphi\,dz$$

where the limits of integration are still to be set up as in Sec. 11 where we set the limits in Cartesian coordinates.

If we use spherical coordinates (see Sec. X.1) the element of volume can be again regarded as being approximately equal to the
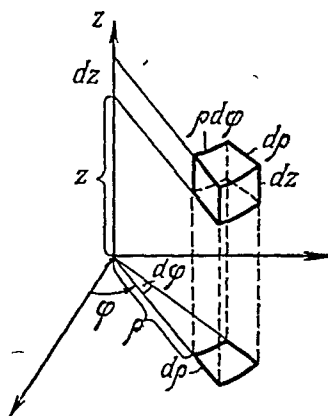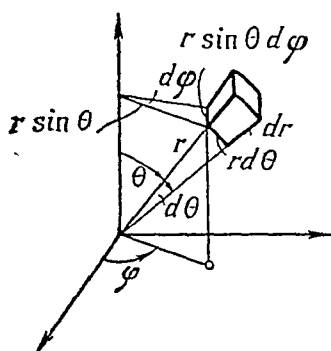
Fig. 316   Fig. 317

volume of the corresponding rectangular parallelepiped (see Fig. 317). In this case the rectangular parallelepiped has the sides dr, $r\,d\theta$ and $r\sin\theta\,d\varphi$ and thus we have

$$d\Omega = dr\cdot r\,d\theta\cdot r\sin\theta\,d\varphi = r^2\sin\theta\,dr\,d\theta\,d\varphi$$

Consequently, integral (31) takes the form

$$I = \iiint\limits_{(\Omega)} ur^2\sin\theta\,dr\,d\theta\,d\varphi \tag{32}$$

The limits of integration are set up in a particularly simple manner in these coordinate systems (and also in other systems) when the boundary of the region (Ω) consists of coordinate surfaces because

in such a case not only the limits of the outer integration are constant but of the first and second integrations as well.

As an example, let us consider the problem of determining the position of the geometrical centre of gravity of a solid having the form of a hemisphere of radius R. To do this we place the hemisphere as it is shown in Fig. 318. Then the symmetry implies that the centre
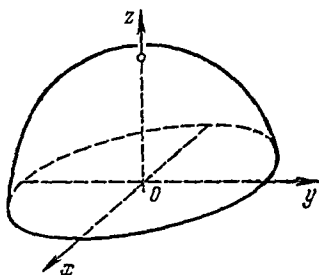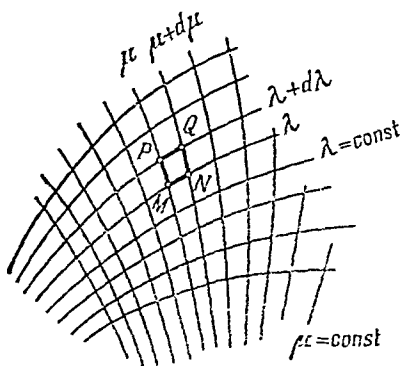
Fig. 318   Fig. 319

of gravity will lie on the z-axis. Taking advantage of formula (9), passing to spherical coordinates by means of formula (32) and taking into account that $z = r \cos \theta$ we obtain the following expression:

$$z_c = \frac{1}{\frac{2}{3}\pi R^3} \int\int\int_{(\Omega)} r \cos \theta \cdot r^2 \cdot \sin \theta \, dr \, d\theta \, d\varphi =$$

$$= \frac{3}{2\pi R^3} \int_0^{2\pi} d\varphi \int_0^{\frac{\pi}{2}} d\theta \int_0^R r^3 \sin \theta \cos \theta \, dr = \frac{3}{8} R$$

(Check up the calculations!)

14. Curvilinear Coordinates in Plane. Besides Cartesian and polar coordinates, we can introduce many other coordinate systems in plane. Their common feature is that the points of the plane are always characterized by two coordinates (see Sec. X.2).

We now consider a general coordinate system λ, μ whose coordinate curves $\lambda = $ const and $\mu = $ const are depicted in Fig. 319. If these curves are drawn sufficiently close to one another the plane is divided into small figures (cells) which can be regarded as parallelograms to within infinitesimals of higher order.

Let the curves $\lambda = $ const be drawn with the interval $d\lambda$ and the curves $\mu = $ const with the interval $d\mu$. We denote the sides of one

of the small parallelograms by $ds_\lambda = MP$ and $ds_\mu = MN$ (see Fig. 319). If we neglect the infinitesimals of higher order we can consider these sides to be directly proportional to $d\lambda$ and $d\mu$, i.e.

$$ds_\lambda = l_\lambda\, d\lambda, \quad ds_\mu = l_\mu\, d\mu \tag{33}$$

The quantities l_λ and l_μ are called **Lamé's coefficients** after G. Lamé (1795-1870), a French mathematician and engineer. The coefficients make it possible to compute linear sizes in a curvilinear coordinate system. For a given coordinate system, Lamé's coefficients can have different values at different points of the plane in the general case. For instance, Fig. 319 indicates that these values are smaller in the lower part of the plane than in the upper (why?). If it is necessary to calculate the arc length of a finite portion of a coordinate curve we must integrate the corresponding relation (33).

Let us introduce the radius-vector $\mathbf{r} = \mathbf{r}\,(\lambda,\ \mu) = \overrightarrow{OM}$ of a variable point M of the plane drawn from a fixed point O. Then the sides $\overrightarrow{MP}$ and $\overrightarrow{MN}$ of the elementary parallelogram depicted in Fig. 319 are equal to

$$\partial_\lambda \mathbf{r} = \mathbf{r}'_\lambda\, d\lambda \quad \text{and} \quad \partial_\mu \mathbf{r} = \mathbf{r}'_\mu\, d\mu$$

to within infinitesimals of higher order since these increments of the radius-vector are due to the variation of only one of the coordinates. It follows that $|\,\partial_\lambda \mathbf{r}\,| = |\,\mathbf{r}'_\lambda\,|\, d\lambda$. But, according to Sec. VII.23, we have $|\,d\mathbf{r}\,| = ds$, and therefore $|\,\partial_\lambda \mathbf{r}\,| = ds_\lambda$. Hence, taking advantage of (33), we derive

$$l_\lambda = |\,\mathbf{r}'_\lambda\,|, \quad \text{and similarly} \quad l_\mu = |\,\mathbf{r}'_\mu\,|$$

If besides the curvilinear coordinates $\lambda,\ \mu$ we introduce Cartesian coordinates x, y whose origin is placed at the same point O we shall have $\mathbf{r} = x\mathbf{i} + y\mathbf{j}$ (see Sec. VII.9), and consequently

$$l_\lambda = \left|\frac{\partial x}{\partial \lambda}\,\mathbf{i} + \frac{\partial y}{\partial \lambda}\,\mathbf{j}\right| = \sqrt{\left(\frac{\partial x}{\partial \lambda}\right)^2 + \left(\frac{\partial y}{\partial \lambda}\right)^2}$$

and

$$l_\mu = \left|\frac{\partial x}{\partial \mu}\,\mathbf{i} + \frac{\partial y}{\partial \mu}\,\mathbf{j}\right| = \sqrt{\left(\frac{\partial x}{\partial \mu}\right)^2 + \left(\frac{\partial y}{\partial \mu}\right)^2}$$

Let the reader derive the formulas $l_\rho = 1$ and $l_\varphi = \rho$ for a polar coordinate system whose origin coincides with the origin of the Cartesian system x, y first on the basis of formulas $x = \rho \cos \varphi$, $y = \rho \sin \varphi$ and then directly by taking advantage of formulas (33).

The area $d\sigma$ of any elementary parallelogram shown in Fig. 319 is proportional both to $d\lambda$ and $d\mu$, i.e.

$$d\sigma = k\, d\lambda\, d\mu \tag{34}$$

where k is a coefficient which can take on different values at different points in the general case. Applying the formula of the area of a parallelogram we obtain

$$k = \frac{d\sigma}{d\lambda \, d\mu} = \frac{ds_\lambda \, ds_\mu \sin \alpha}{d\lambda \, d\mu} = l_\lambda l_\mu \sin \alpha \qquad (35)$$

where α is the angle between the corresponding coordinate curves.

In particular, for an **orthogonal coordinate system**, i.e. a system whose coordinate curves intersect at right angles, we have

$$k = l_\lambda l_\mu \qquad (36)$$

In the general case we can deduce from formulas (34) and (VII.21) the formula

$$k = \left| \begin{vmatrix} \dfrac{\partial x}{\partial \lambda} & \dfrac{\partial y}{\partial \lambda} \\ \dfrac{\partial x}{\partial \mu} & \dfrac{\partial y}{\partial \mu} \end{vmatrix} \right| = \left| \frac{D(x, y)}{D(\lambda, \mu)} \right| \qquad (37)$$

(see Sec. IX.13 on the last notation). In deducing the formula we apply the property of determinants (property 7 in Sec. VI.2) according to which a determinant does not change its value when being transposed.

The same result can be obtained if we note that the coefficient k in formula (34) characterizes the change of areas under the mapping of the λ, μ-plane on the x, y-plane defined by the formulas $x = x(\lambda, \mu)$, $y = y(\lambda, \mu)$. By Sec. XI.14, this coefficient is equal to the absolute value of the corresponding Jacobian, i.e. of the determinant $\dfrac{D(x, y)}{D(\lambda, \mu)}$ entering into (37).

[Let the reader deduce the formula $k = \rho$ for a polar coordinate system taking advantage of formula (37). Do the same on the basis of formula (36). Find Lamé's coefficients and the coefficient k for a Cartesian coordinate system.]

If we take an integral of the form

$$I = \int\limits_{(\sigma)} u \, d\sigma$$

taken over a finite plane region (σ) we obtain, on the basis of formula (34), the expression

$$I = \int\limits_{(\sigma)} \int u k \, d\lambda \, d\mu \qquad (38)$$

where the limits of integration must be set up by analogy with integrals (23), (24) and (30). The simplest case for setting up the limits of integration is when the region in question is bounded by

a contour consisting of arcs of the corresponding coordinate curves (why is it so?). [Let the reader verify that formula (38) turns into formula (30) in the case of polar coordinates.]

15. Curvilinear Coordinates in Space. General curvilinear coordinates λ, μ, ν in space are considered in a similar way. The surfaces $\lambda = \text{const}$, $\mu = \text{const}$ and $\nu = \text{const}$ form three families of coordinate surfaces whose *intersections generate three families of coordinate curves.* The coordinate surfaces corresponding to the values λ, $\lambda + d\lambda$; μ, $\mu + d\mu$ and ν, $\nu + d\nu$ of the coordinates bound an elementary volume (cell) in space which can be regarded as a parallelepiped to within infinitesimals of higher order (the parallelepiped can be oblique in the general case). Such parallelepipeds are shown in Figs. 316 and 317 for the concrete cases of cylindrical and spherical coordinates (in these cases the parallelepipeds are rectangular). The length of one of the edges of the infinitesimal parallelepiped is equal to

$$ds_\lambda = |\partial_\lambda \mathbf{r}| = |\mathbf{r}'_\lambda| \, d\lambda = l_\lambda \, d\lambda$$

(to within infinitesimals of higher order) where

$$l_\lambda = |\mathbf{r}'_\lambda| = \sqrt{\left(\frac{\partial x}{\partial \lambda}\right)^2 + \left(\frac{\partial y}{\partial \lambda}\right)^2 + \left(\frac{\partial z}{\partial \lambda}\right)^2}$$

is a Lamé coefficient. The lengths of the other two edges of the parallelepiped are expressed similarly. The volume of the parallelepiped is equal to $d\Omega = k \, d\lambda \, d\mu \, d\nu$ where k is a coefficient which can take on different values at different points in space. Therefore the corresponding change of variables in a triple integral is performed according to the formula

$$\int_{(\Omega)} u \, d\Omega = \int\int_{(\Omega)}\int [uk \, d\lambda] d\mu \, d\nu \tag{39}$$

In the case of an orthogonal coordinate system we have $k = l_\lambda l_\mu l_\nu$. In the general case the coefficient k can be found on the basis of the geometric meaning of a triple scalar product of vectors (see Sec. VII.15):

$$k = \frac{d\Omega}{d\lambda \, d\mu \, d\nu} = \frac{|(\partial_\lambda \mathbf{r} \times \partial_\mu \mathbf{r}) \cdot \partial_\nu \mathbf{r}|}{d\lambda \, d\mu \, d\nu} = \frac{|(\mathbf{r}'_\lambda \, d\lambda \times \mathbf{r}'_\mu \, d\mu) \cdot \mathbf{r}'_\nu \, d\nu|}{d\lambda \, d\mu \, d\nu} =$$

$$= |(\mathbf{r}'_\lambda \times \mathbf{r}'_\mu) \cdot \mathbf{r}'_\nu| = \left| \begin{matrix} \dfrac{\partial x}{\partial \lambda} & \dfrac{\partial y}{\partial \lambda} & \dfrac{\partial z}{\partial \lambda} \\[2mm] \dfrac{\partial x}{\partial \mu} & \dfrac{\partial y}{\partial \mu} & \dfrac{\partial z}{\partial \mu} \\[2mm] \dfrac{\partial x}{\partial \nu} & \dfrac{\partial y}{\partial \nu} & \dfrac{\partial z}{\partial \nu} \end{matrix} \right| = \left| \frac{D(x, y, z)}{D(\lambda, \mu, \nu)} \right|$$

39*

(Let the reader calculate the coefficients l_λ, l_μ, l_ν and k for Cartesian, cylindrical and spherical coordinates.)

16. Coordinates on a Surface. It is possible to introduce a coordinate system on an arbitrary surface (see Sec. X.6 and Fig. 211). Let us denote the coordinates by the letters λ and μ. After a manner of Sec. 14, we can express, to within infinitesimals of higher order, the sides and the area of an infinitesimal parallelogram bounded by the coordinate curves corresponding to the values λ, $\lambda + d\lambda$ and μ, $\mu + d\mu$ of the coordinates. Indeed,

$$ds_\lambda = |\,\partial_\lambda \mathbf{r}\,| = l_\lambda\, d\lambda$$

where

$$l_\lambda = |\,\mathbf{r}'_\lambda\,| = \sqrt{\left(\frac{\partial x}{\partial \lambda}\right)^2 + \left(\frac{\partial y}{\partial \lambda}\right)^2 + \left(\frac{\partial z}{\partial \lambda}\right)^2}$$

(ds_μ is expressed similarly). Thus, we have $d\sigma = k\, d\lambda\, d\mu$ where $k = l_\lambda l_\mu$ for an orthogonal coordinate system and

$$k = |\,\mathbf{r}'_\lambda \times \mathbf{r}'_\mu\,| = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial x}{\partial \lambda} & \dfrac{\partial y}{\partial \lambda} & \dfrac{\partial z}{\partial \lambda} \\ \dfrac{\partial x}{\partial \mu} & \dfrac{\partial y}{\partial \mu} & \dfrac{\partial z}{\partial \mu} \end{vmatrix} =$$

$$= \sqrt{\left(\frac{\partial y}{\partial \lambda}\frac{\partial z}{\partial \mu} - \frac{\partial z}{\partial \lambda}\frac{\partial y}{\partial \mu}\right)^2 + \left(\frac{\partial x}{\partial \lambda}\frac{\partial z}{\partial \mu} - \frac{\partial z}{\partial \lambda}\frac{\partial x}{\partial \mu}\right)^2 + \left(\frac{\partial x}{\partial \lambda}\frac{\partial y}{\partial \mu} - \frac{\partial y}{\partial \lambda}\frac{\partial x}{\partial \mu}\right)^2}$$

in the general case. The transformation of a surface integral to the variables λ and μ is performed according to formula (38).

As an example, let us consider the surface of a sphere of fixed radius R. The spherical coordinates φ, θ considered on the sphere are the example of curvilinear coordinates on a surface. These coordinates are expressed as ordinary spherical coordinates in space with a fixed value $r = R$ of the radius. This is an orthogonal system, and Fig. 317 directly implies that

$$ds_\varphi = R \sin \theta\, d\varphi, \quad ds_\theta = R\, d\theta$$

i.e.

$$l_\varphi = R \sin \theta, \quad l_\theta = R$$

The coefficient k is expressed as $k = l_\varphi l_\theta = R^2 \sin \theta$ in this case. Consequently, an integral taken over a portion (σ) of the sphere (which can coincide with the whole sphere) is computed by the

formula

$$\int\limits_{(\sigma)} u \, d\sigma = R^2 \int\limits_{(\sigma)} \int u \sin \theta \, d\varphi \, d\theta \qquad (40)$$

As an example, let us determine the force of attraction between a material point of mass m and the surface (σ) of the whole material sphere of radius R with a constant surface density of mass ρ. Because of the symmetry, we can, without loss of generality, limit ourselves to the disposition shown in Fig. 320. Every surface element $d\sigma$ attracts the mass m with the force dF which can be found on the basis of Newton's law of gravitation:

(6)

$$|dF| = \varkappa \frac{m\rho \, d\sigma}{l^2} = \varkappa \frac{m\rho}{R^2 + h^2 - 2Rh \cos \theta} \, d\sigma \qquad (41)$$

where $\varkappa$ is the constant of gravitation.

When summing up these elementary forces we must add together the projections of the forces on the coordinate axes but not the absolute values of the forces because they have different directions. The symmetry implies that the resultant force is directed along the z-axis and therefore we have to add together the projections of all the elementary forces on the z-axis:
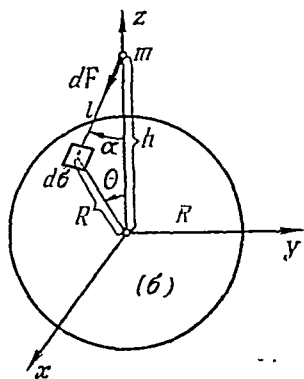
Fig. 320

$$F = \int\limits_{(\sigma)} (dF)_z = \int\limits_{(\sigma)} |dF| \cos \alpha =$$

$$= \int\limits_{(\sigma)} \varkappa \frac{m\rho}{R^2 + h^2 - 2Rh \cos \theta} \, d\sigma \frac{l^2 + h^2 - R^2}{2hl}$$

(the expression of $\cos \alpha$ has been found on the basis of the cosine law which can be written as $R^2 = l^2 + h^2 - 2lh \cos \alpha$ in this case).

Substituting $l = \sqrt{R^2 + h^2 - 2Rh \cos \alpha}$ into the last integral and passing to spherical coordinates according to formula (40) we obtain

$$F = \varkappa m\rho \int\limits_{(\sigma)} \frac{h - R \cos \theta}{(R^2 + h^2 - 2Rh \cos \theta)^{\frac{3}{2}}} \, d\sigma =$$

$$= \varkappa m\rho \int\limits_{0}^{\pi} d\theta \int\limits_{0}^{2\pi} \frac{h - R \cos \theta}{(R^2 + h^2 - 2Rh \cos \theta)^{\frac{3}{2}}} R^2 \sin \theta \, d\varphi =$$

$$= R^2 \varkappa m \rho \int\limits_{0}^{\pi} \frac{h - R \cos \theta}{(R^2 + h^2 - 2Rh \cos \theta)^{\frac{3}{2}}} \sin \theta \, d\theta \int\limits_{0}^{2\pi} d\varphi =$$

$$= - 2\pi R^2 \varkappa m \rho \int\limits_{0}^{\pi} \frac{h - R \cos \theta}{(R^2 + h^2 - 2Rh \cos \theta)^{\frac{3}{2}}} \, d \cos \theta =$$

$$= 2\pi R^2 \varkappa m \rho \int\limits_{-1}^{1} \frac{h - Rt}{(R^2 + h^2 - 2Rht)^{\frac{3}{2}}} \, dt$$

Performing the substitution $R^2 + h^2 - 2Rht = l^2$ $(l > 0)$, $-2Rh \, dt = 2l \, dl$ we finally deduce:

$$F = 2\pi R^2 \varkappa m \rho \int\limits_{R-h}^{|R-h|} \frac{h + \frac{l^2 - R^2 - h^2}{2h}}{l^3} \left(-\frac{2l \, dl}{2Rh} \right) =$$

$$= \frac{\pi R \varkappa m \rho}{h^2} \int\limits_{|R-h|}^{R+h} \left(1 + \frac{h^2 - R^2}{l^2} \right) \! dl =$$

$$= \frac{\pi R \varkappa m \rho}{h^2} \left[(R + h) - |R - h| + (h^2 - R^2) \left(\frac{1}{|R-h|} - \frac{1}{R+h} \right) \right] \quad (42)$$

If $h > R$ we have $|R - h| = h - R$. Substituting this expression into formula (42) we obtain (check it up!) the formula

$$F = \varkappa \frac{m 4\pi R^2 \rho}{h^2} = \varkappa \frac{mM}{h^2} \quad (h > R)$$

where M is the total mass of the sphere. If $h < R$ we have $|R - h| = R - h$, and thus we similarly find

$$F = 0 \quad (h < R)$$

Hence, a homogeneous sphere attracts material points lying outside it, as if the total mass of the sphere were concentrated at its centre, and does not attract points lying inside it. Now let us consider a material solid of spherical form with a spherically symmetric mass distribution (that is the density depends only on the distance from the centre). Such a solid can be thought of as consisting of spherical layers of infinitesimal width bounded by concentric spheres. Each layer can be regarded as a material surface to which the above result can be applied. Thus, we conclude that the spherical solid attracts a point lying outside it as if the whole mass were concentrated at its centre. Similarly, a point lying inside the solid is attracted only by the portion of the solid which lies closer to the centre than the point.

§ 5. Other Types of Multiple Integrals

17. Improper Integrals. The theory of improper multiple integrals is similar to that of one-dimensional integrals (see § XIV.4). We begin with the integral

$$I = \int\limits_{(\Omega)} u \, d\Omega \tag{43}$$

in which the integrand u is a finite function and the domain of integration is unbounded (infinite). It is defined as the limit

$$\int\limits_{(\Omega)} u \, d\Omega = \lim\limits_{(\Omega') \to (\Omega)} \int\limits_{(\Omega')} u \, d\Omega' \tag{44}$$

where the domain (Ω') on the right-hand side is finite. This domain expands in the limiting process and exhausts the whole domain (Ω) in the limit (see Fig. 321). If limit (44) exists, is finite and independent of a particular way in which the domain (Ω') expands integral (43) is said to be convergent. If otherwise the integral is referred to as being divergent. If limit (44) equals infinity we write $\int\limits_{(\Omega)} u \, d\Omega = \infty$. If $u \geqslant 0$ integral (44) either converges or diverges and $\int\limits_{(\Omega)} u \, d\Omega = +\infty$ (i.e. it is divergent to infinity).

Boundary of the domain (Ω)

Boundary of the domain (Ω')

Fig. 321

In such a case we can set up the limits of integration in any coordinate system (convenient for the computation) by the rules given in §§ 3, 4. If the result of the substitution of the limits is finite the integral is convergent and if otherwise the integral is divergent. The comparison tests [see (XIV.49) and (XIV.50)] remain valid in this case. In applying the comparison tests we can use integrals (XIV.51) and some other integrals. For instance, in the case when (Ω) is the whole plane we often perform the comparison with an integral of the function r^{-p} where $r = \sqrt{x^2 + y^2}$ is the length of the radius-vector. It is only the behaviour of the integrand for large values of r that is essential for the convergence of an integral of this type, and therefore we must investigate the integral
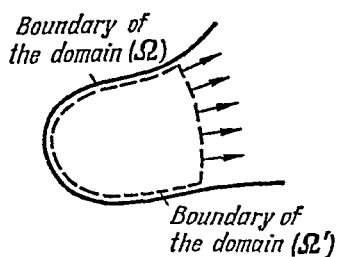
$$\iint\limits_{(r > r_0)} r^{-p} \, dx \, dy = \int\limits_0^{2\pi} d\varphi \int\limits_{r_0}^{\infty} r^{-p} r \, dr = 2\pi \int\limits_{r_0}^{\infty} \frac{1}{r^{p-1}} \, dr \quad (r_0 > 0)$$

(we have passed to polar coordinates here). According to Sec. XIV.15 [see formula (XIV.51)], the integral is finite for $p > 2$ and infinite for $p \leqslant 2$. Similarly, in the three-dimensional space x, y, z the integral of $r^{-p} = (\sqrt{x^2 + y^2 + z^2})^{-p}$ (taken over the whole space with a sphere of radius $r_0 > 0$ and centre at the origin of coordinates cut out) converges only for $p > 3$.

As an example of an application of improper multiple integrals, let us deduce formula (XIV.70). We take the integral

$$I = \int\limits_0^\infty \int\limits_0^\infty x^{p-1} y^{p+q-1} e^{-(x+1)\,y}\, dx\, dy \quad (p > 0, \quad q > 0)$$

whose domain of integration is the first quadrant of the x, y-plane. The integrand being a positive function, we can perform the integration in any order. This yields the following results:

$$(1) \quad I = \int\limits_0^\infty dy \int\limits_0^\infty x^{p-1} y^{p+q-1} e^{-(x+1)\,v}\, dx =$$

$$= \int\limits_0^\infty dy \int\limits_0^\infty \left(\frac{s}{y}\right)^{p-1} y^{p+q-1} e^{-\left(\frac{s}{y}+1\right)y}\, \frac{ds}{y}$$

$\left(\text{we have made the substitution } x = \dfrac{s}{y}\right)$. Calculating we find:

$$I = \int\limits_0^\infty s^{p-1} e^{-s}\, ds \cdot \int\limits_0^\infty y^{q-1} e^{-y}\, dy = \Gamma(p)\, \Gamma(q)$$

$$(2) \quad I = \int\limits_0^\infty dx \int\limits_0^\infty x^{p-1} y^{p+q-1} e^{-(x+1)\,v}\, dy =$$

$$= \int\limits_0^\infty dx \int\limits_0^\infty x^{p-1} \left(\frac{t}{x+1}\right)^{p+q-1} e^{-t}\, \frac{dt}{x+1}$$

$\left(\text{we have employed the change of variable } y = \dfrac{t}{x+1}\right)$. Now, applying formula (XIV.73) we deduce:

$$I = \int\limits_0^\infty \frac{x^{p-1}}{(x+1)^{p+q}}\, dx \cdot \int\limits_0^\infty t^{p+q-1} e^{-t}\, dt = B(p, q)\, \Gamma(p+q)$$

Comparing the results we derive the desired formula.

If the function u takes on the values of both signs and

$$\int\limits_{(\Omega)} |u|\, d\Omega < \infty \tag{45}$$

integral (43) converges and is said to be **absolutely convergent.** In evaluating such an integral we can choose any convenient coordinate system and set up the limits of integration. It can also be proved that if condition (45) is violated integral (43) is divergent. In this case limit (44) can depend on the way in which the domain (Ω') expands. Then it can happen that after setting up the limits of integration and evaluating the integral we obtain a finite result in one coordinate system, an infinite result in some other system, a finite result but different from the first one in a third system and a divergence of an oscillating type in a fourth coordinate system (see Sec. XIV.14). Hence, in such a case the possibility of changing variables and inverting the order of integration should be additionally investigated. There are no such problems in evaluating absolutely convergent integrals.

Improper multiple integrals of other types are treated similarly. Namely, if the domain of integration contains a point, a curve or a surface on which the integrand approaches infinity, the singularity (i.e. the point, the curve or the surface) is cut out of the domain together with its neighbourhood and then the boundary of the cut-out portion is contracted to the singularity in an arbitrary way. The limit thus obtained is taken as the value of the improper integral provided this limit is finite and independent of the way of the contraction. If an improper integral of this type of a function which is positive everywhere or positive near its singularities converges the integration can be performed in any coordinate system. The same can be done for arbitrary functions if they are absolutely integrable. When investigating an integral whose integrand has an *isolated singularity,* that is a separate point at which the integrand approaches infinity, we often apply the comparison test using the integral $\iint\limits_{r \leqslant r_0} r^{-p}\, dx\, dy$ in plane and the integral $\iiint\limits_{r \leqslant r_0} r^{-p}\, dx\, dy\, dz$ in space. We can easily verify that the former converges only for $p < 2$ and the latter for $p < 3$. If there is a non-isolated singularity then in investigating the convergence it is convenient to choose a coordinate system in such a way that the singularity should coincide with one of the coordinate curves or surfaces.

18. Integrals Dependent on a Parameter. Let us consider integrals of the form

$$I(\lambda) = \int\limits_{(\Omega)} f(M, \lambda)\, d\Omega$$

where M is a variable point in the domain (Ω) whose coordinates are the variables of integration and λ is a parameter which is kept constant in the process of integration. The theory of such integrals is developed by analogy with Sec. XIV.5. All the basic assertions

proved there remain valid here. There is a difficulty in this case which arises when the domain of integration also depends on the parameter. In such cases we often try to change the variables so that the new domain of integration should become fixed. But, of course, such integrals can also be investigated directly.

For instance, let us consider a triple integral of the form

$$I(\lambda) = \iiint\limits_{(\varphi_\lambda(x,\, y,\, z) \leqslant 0)} f(M)\, d\Omega$$

where the function $\varphi_\lambda(x, y, z)$ depends on the parameter λ and the domain of integration is a region in which $\varphi_\lambda \leqslant 0$. Let it be necessary to compute the derivative $\frac{dI}{d\lambda}$. The expression dI coincides, to within infinitesimals of higher order, with the difference $I(\lambda + d\lambda) - I(\lambda)$ which is equal to an integral of $f(M)$ taken over a thin layer bounded by the surfaces (S_λ) and $(S_{\lambda+d\lambda})$ with the equations $\varphi_\lambda = 0$ and $\varphi_{\lambda+d\lambda} = 0$, respectively. Let us take a point A on (S_λ) and draw the normal to (S_λ) at A. We reckon the distances from A along the normal and consider them positive in the direction to the region where $\varphi_\lambda > 0$, that is in the direction of outer normal to the boundary surface of the domain of integration. We now designate the point of intersection of the normal with the surface $(S_{\lambda-d\lambda})$ by $\overline{A}$. Then the quantity $dn = A\overline{A}$ is equal to the width of the layer at the point A. We have $\varphi_\lambda(A) = 0$ because the point A belongs to (S_λ). We also have $\varphi_{\lambda+d\lambda}(\overline{A}) = 0$ (why?). Now we can write the relation

$$\varphi_{\lambda+d\lambda}(\overline{A}) = \varphi_\lambda(\overline{A}) + \frac{\partial\varphi}{\partial\lambda}\, d\lambda = \varphi_\lambda(A) + |\operatorname{grad}\ \varphi_\lambda|\, dn + \frac{\partial\varphi}{\partial\lambda}\, d\lambda$$

which is accurate to within infinitesimals of higher order. It follows that

$$|\operatorname{grad}\varphi_\lambda|\, dn + \frac{\partial\varphi}{\partial\lambda}\, d\lambda = 0, \quad \text{i.e.} \quad dn = -\frac{\dfrac{\partial\varphi}{\partial\lambda}}{|\operatorname{grad}\varphi_\lambda|}\, d\lambda$$

The element of volume of the layer can be written as $d\Omega = dS\, dn$ where dS is the surface element of (S_λ). Consequently, we have

$$d\Omega = dS\, dn = -\frac{\dfrac{\partial\varphi}{\partial\lambda}}{|\operatorname{grad}\varphi|}\, dS\, d\lambda$$

The quantity dn can be positive or negative here. Its sign indicates whether the volume of the layer is added to the volume of the original domain (corresponding to the value λ of the parameter) or sub-

tracted from it. It follows that

$$dI = \int \int \int f \cdot \left(-\frac{\frac{\partial \varphi}{\partial \lambda}}{|\operatorname{grad} \varphi|} \, dS \, d\lambda \right), \quad \text{i.e.} \quad \frac{dI}{d\lambda} =$$

$$= -\int \int \int_{(\varphi_\lambda = 0)} f(M) \frac{\frac{\partial \varphi}{\partial \lambda}}{|\operatorname{grad} \varphi|} \, dS$$

An integral can depend on several parameters. The coordinates of a point N varying within a certain region can play the role of such parameters. An integral of this type is written as

$$I(N) = \int_{(\Omega)} f(M, N) \, d\Omega_M$$

where the symbol $d\Omega_M$ indicates the fact that it is the moving point M whose coordinates are the variables of integration (see Sec. 2). The point N which can occupy different positions is kept fixed in the process of integration. Its coordinates are the parameters. The basic properties of integrals dependent on a parameter (see Sec. XIV.5) are easily extended to these integrals.

An integral can be integrated with respect to a parameter on which it depends. This results in a multiple integral of higher order. For example, let us consider the problem of calculating the force F of attraction between two material bodies (Ω_1) and (Ω_2) whose densities ρ_1 and ρ_2 can be variable in the general case. Let us take two elements of volume $d\Omega_1$ and $d\Omega_2$ placed at some points M_1 and M_2 belonging to (Ω_1) and (Ω_2), respectively. The force with which the element $d\Omega_1$ attracts the element $d\Omega_2$ can be expressed on the basis of Newton's law of gravitation:

$$d \, d\mathbf{F} = \varkappa \frac{\rho_1 \, d\Omega_1 \cdot \rho_2 \, d\Omega_2}{|\overrightarrow{M_2 M_1}|^2} (\overrightarrow{M_2 M_1})^0 = \varkappa \frac{\rho_1 \rho_2 \overrightarrow{M_2 M_1}}{|\overrightarrow{M_2 M_1}|^3} \, d\Omega_1 \, d\Omega_2$$

where $(\overrightarrow{M_2 M_1})^0$ is the unit vector in the direction of the vector $\overrightarrow{M_2 M_1}$. Now, integrating over (Ω_1), we obtain the force with which the whole body (Ω_1) attracts the element $d\Omega_2$:

$$d\mathbf{F} = \varkappa \left(\int_{(\Omega_1)} \frac{\rho_1 \overrightarrow{M_2 M_1}}{|\overrightarrow{M_2 M_1}|^3} \, d\Omega_1 \right) \rho_2 \, d\Omega_2$$

In the above integral the integration is performed with respect to the coordinates of the variable point M_1 running throughout the domain (Ω_1) when the point M_2 is arbitrarily fixed in the domain (Ω_2), the coordinates of M_2 being parameters. To obtain the resultant force of attraction we must additionally integrate with respect

to the coordinates of M_2:

$$\mathbf{F} = \varkappa \int\limits_{(\Omega_2)} \rho_2 \, d\Omega_2 \int\limits_{(\Omega_1)} \frac{\rho_1 \overrightarrow{M_2 M_1}}{|\overrightarrow{M_2 M_1}|^3} \, d\Omega_1$$

If we set up the limits of integration we shall obtain a six-fold iterated integral. For instance, in Cartesian coordinates we shall have

$$\mathbf{F} = \varkappa \iiint\limits_{(\Omega_2)} \rho_2 \, (x_2, \, y_2, \, z_2) \, dx_2 \, dy_2 \, dz_2 \, \times$$

$$\times \iiint\limits_{(\Omega_1)} \frac{\rho_1 \, (x_1, \, y_1, \, z_1) \, [(x_1 - x_2) \, \mathbf{i} + (y_1 - y_2) \, \mathbf{j} + (z_1 - z_2) \, \mathbf{k}]}{[(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2]^{3/2}} \, dx_1 \, dy_1 \, dz_1$$

The limits of integration which are to be set up in the above integral depend on the form of the domains (Ω_1) and (Ω_2).

19. Integrals with Respect to Measure. Generalized Functions. For definiteness, let us consider triple integrals. In Sec. 1, the volume of a domain in space was called its measure. But this is only the simplest example of a measure which is referred to as the *Lebesgue measure* (after the French mathematician H. Lebesgue, 1875-1941, who developed the general theory of the measure). It is also possible to introduce other measures with respect to which integration can be performed.

We begin with an example. Let a mass m be distributed in space (see Sec. 6). By analogy with Sec. 18, we can readily deduce the expression

$$\mathbf{F} = \varkappa m_0 \int \frac{\overrightarrow{NM}}{|\overrightarrow{NM}|^3} \, dm \tag{46}$$

of the force with which the distributed mass m acts upon a mass point m_0 placed at the point N. Here M is a variable point in space whose coordinates are the variables of integration and the integration extends over the whole region of space in which the mass m is distributed.

If the mass m is distributed continuously, that is if the mass of every surface, curve or point is equal to zero, we can introduce the density ρ (see Sec. 6). and then integral (46) can be reduced to an ordinary triple integral

$$\int \frac{\overrightarrow{NM}}{|\overrightarrow{NM}|^3} \, \rho \, d\Omega \tag{47}$$

But we sometimes deal with a mass which is concentrated on separate surfaces, curves or at points. Then the ordinary notion of density cannot be applied and hence it is impossible to pass to

integral (47). In such cases we have to consider integral (46) as an *integral with respect to the measure m.*

The general definition of a measure in space is analogous to the basic definition given in Sec. 7. To every part (Ω) of space which can be mentally isolated from it (i.e. to each solid, surface, curve or point etc.), a certain value $\mu_{(\Omega)}$ of the measure μ must correspond. The condition of additivity is also imposed on μ. The measure of a surface, curve or point may not be equal to zero in the general case. We usually deal with non-negative measures $\mu \geqslant 0$ but it is sometimes expedient to consider *signed measures* (also called *charges*) which can assume values of either sign. In such cases the measure can be thought of not as a mass but as an electric charge which can be positive or negative. A measure can be defined not only in space but also on a surface or curve.

The definition of an integral with respect to measure is similar to the ordinary definition given in Sec. 2 (integrals with respect to measure are also called the *Stieltjes integrals*). Let us take a domain (Ω) in space. If a measure μ and a function $u\ (M)$ [where M is a variable point running over (Ω)] are defined in the domain (Ω) the integral of u with respect to the measure μ is defined as

$$\int_{(\Omega)} u\, d\mu = \lim \sum_{h=1}^{n} u\ (M_h)\, \mu_{(\Delta\Omega_h)} \tag{48}$$

where the meaning of the notation is obvious. Such an integral always exists provided the function u is finite in (Ω) and the measure of the whole domain (Ω) is also finite (if μ is not non-negative we must additionally impose the condition that $\int_{(\Omega)} |\, d\mu\,| < \infty$ which means that the *positive* and *negative variations* (parts) of μ should also be finite). If the function u is discontinuous the form of integral sums used in (48) must also be specified but we shall not discuss this question at length here. Improper integrals with respect to measure are defined after a manner of Sec. 17. The properties of integral (48) are similar to those discussed in Sec. 3. When we speak about the properties related to integrating inequalities we must additionally impose the condition $\mu \geqslant 0$.

If the measure of every surface, curve or point is equal to zero it is possible to pass to an ordinary triple integral taken over a volume:

$$\int_{(\Omega)} u\, d\mu = \int_{(\Omega)} u\, \frac{d\mu}{d\Omega}\, d\Omega = \int_{(\Omega)} u\rho\, d\Omega \quad \left(\rho = \frac{d\mu}{d\Omega}\right) \tag{49}$$

This transition can be performed for any measure but in the general case ρ will be a **generalized function.**

The simplest generalized function in space is the delta function

$$\delta (x - a)\, \delta (y - b)\, \delta (z - c) \qquad (50)$$

(see Sec. XIV.25) which describes the volume density of unit mass placed at the point $(a,\ b,\ c)$. The function $\delta (y - b)\, \delta (z - c)$ describes the volume density of a mass uniformly distributed along the straight line $y = b$, $z = c$ with unit linear density. The function $\delta (z - c)$ is the volume density of a mass uniformly distributed over the plane $z = c$ with unit areal (surface) density. Using these functions and some other generalized functions (in particular, delta functions depending on curvilinear coordinates) we can perform transformation (49) in the general case.

The properties of the generalized functions of several variables are similar to those of functions of one variable (see Sec. XIV.27). Generalized function (50) can be applied to constructing an influence function (Green's function; see Sec. XIV.26) of the form

$$G (M,\ N) = G (x,\ y,\ z,\ \xi,\ \eta,\ \zeta)$$

where $(x,\ y,\ z)$ are the coordinates of the point M (point of observation) and $(\xi,\ \eta,\ \zeta)$ are the coordinates of the point N at which the source (producing the corresponding action) is placed. When investigating processes developing in time we also use the delta function

$$\delta (x - a)\, \delta (y - b)\, \delta (z - c)\, \delta (t - \tau)$$

which yields an influence function of the form $G (M,\ t,\ N,\ \tau)$.

20. **Multiple Integrals of Higher Order.** A measure can also be defined in a k-dimensional space or, as we say, in a *k-dimensional manifold* (see Sec. X.2). The definition of an integral of form (48) and its basic properties remain unchanged in this case. To pass to a repeated integral we must introduce generalized coordinates $t_1, t_2, \ldots \ldots, t_k$ in the manifold (see Sec. X.2), express the integrand as a function of the form $u = u (t_1, \ldots, t_k)$ and find the density $\rho (t_1, \ldots, t_k)$ which defines the element of measure $d\mu = \rho (t_1, \ldots \ldots, t_k)\, dt_1\, dt_2\, \ldots\, dt_k$ corresponding to an infinitesimal generalized k-dimensional parallelepiped placed at the variable point $(t_1, \ldots, t_k)$ and bounded by the corresponding "coordinate surfaces" [which are $(k - 1)$-dimensional submanifolds in the general case]. Then integral (48) takes the form

$$\int\limits_{(\Omega)} u\, d\mu = \underbrace{\int \int \ldots \int}_{k \text{ times}} u (t_1, t_2, \ldots, t_k)\, \rho (t_1, t_2, \ldots, t_k)\, dt_1 \ldots dt_k \qquad (51)$$

where the limits of integration must be set up on the right-hand side according to the ranges of variation of the coordinates $t_1, \ldots, t_k$.

The density ρ entering in formula (51) is understood as an ordinary function if the measure of every submanifold of dimension

$s < k$ (which can be defined by means of one or more equations connecting the coordinates $t_1, \ldots, t_k$) is equal to zero. In particular, this is the case if the density is finite everywhere.

If otherwise, ρ should be understood as a generalized function (see Sec. 19).

If the notion of a volume (hypervolume) is introduced in the space under consideration we can perform the integration with respect to the volume which is a particular case of a measure. To do this we must know the expression

$$d\Omega = h\,(t_1, \ldots, t_k)\, dt_1\, dt_2 \ldots dt_k \tag{52}$$

of the volume of an infinitesimal generalized parallelepiped bounded by the corresponding coordinate surfaces (submanifolds). Then the integral $\int\limits_{(\Omega)} u\, d\Omega$ can be transformed by analogy with (51).

The notion of an integral (with respect to a measure or hypervolume) over a domain belonging to any submanifold of lower dimension lying in the initial k-dimensional manifold is introduced in a similar way. In the ordinary three-dimensional space we can consider line integrals, surface integrals and triple integrals but in a k-dimensional space there are k different types of integral (what are these types?).

For the k-dimensional Cartesian space E_k (see Sec. VII.18) we put $h \equiv 1$ in formula (52), i.e. we take the volume of unit k-dimensional hypercube with unit sides as the unit measure of hypervolume. Integrals of lower order in this space are defined under the convention that the p-dimensional volume ($1 \leqslant p < k$) of a p-dimensional rectangular parallelepiped (finite or infinitesimal) is equal to the product of the lengths of its sides (this is the Lebesgue measure).

By analogy with Sec. XIV.23, we can consider integrals with respect to coordinates taken over a p-dimensional manifold (S) ($1 \leqslant p < k$) lying in E_k. But (S) must be *orientable* in this case. This is a new notion which cannot be easily visualized for $p > 1$, and we are going to discuss it at length here.

First of all we shall introduce the notion of a p-dimensional tetrahedron. By definition, a one-dimensional tetrahedron is a line segment, a two-dimensional tetrahedron is a triangle and a three-dimensional one is a triangular pyramid. To obtain a four-dimensional tetrahedron we take a three-dimensional tetrahedron lying in a three-dimensional space which is considered to be a subspace of some space of dimension $s > 3$. Then we take a point belonging to the s-dimensional space which does not belong to the three-dimensional subspace and connect this point by line segments with all the points of the three-dimensional tetrahedron. The totality of all the points belonging to these line segments is a four-dimensional

tetrahedron. The tetrahedrons of higher dimensions are constructed similarly. Now take a p-dimensional tetrahedron with vertices A_1, A_2, ..., A_{p+1}. Its orientation is specified by enumerating these vertices in a certain order. It is assumed that the permutation of any two vertices changes the orientation to the opposite one. For instance, if we take a three-dimensional tetrahedron with vertices A, B, C, D the combinations $ABCD$ and $DBAC$ define the same orientation whereas the combination $CBAD$ defines the opposite orientation. Every tetrahedron can be oriented in two different ways.

If we take an arbitrary small p-dimensional tetrahedron on a p-dimensional manifold (S), choose a certain orientation for it and then make it run over the manifold, the original orientation
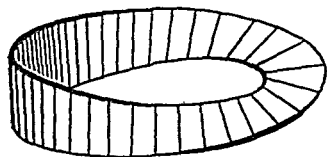
Fig. 322

of the tetrahedron will induce the orientation of all small p-dimensional tetrahedrons on (S). Then we say that (S) has been oriented. For $p = 1$ a manifold of the type (S) is a curve, and the above method of orientation is equivalent to specifying a certain direction on it. For $p = 2$ such a manifold (S) is a two-dimensional surface, and its orientation is equivalent to specifying the direction of describing the contour of any small region on (S). If (S) consists of several disjoint portions they can be oriented independently.

It should be taken into account that in the case $p \geqslant 2$ some manifolds cannot be oriented. The so-called **Möbius strip** (see Fig. 322) discovered in 1858 by the German geometer A. F. Möbius (1790-1868) is the simplest example of a non-orientable two-dimensional surface.

A p-fold multiple integral with respect to coordinates taken over an oriented p-dimensional manifold (S) lying in $E_{(k)}$ is defined as

$$\int \ldots \int_{(S)} u(t_1, \ldots, t_k)\, dt_{m_1} dt_{m_2} \ldots dt_{m_p} = \lim \sum_{k=1}^{n} u(M_h)\, \Delta S_k' \quad (53)$$

where the summation on the right-hand side is extended over all small tetrahedrons (ΔS_h) into which (S) is divided, the orientation of the tetrahedrons being coherent with the orientation of (S). The symbol $\Delta S_h'$ ($k = 1, \ldots, n$) designates the p-dimensional volume of the projection $(\Delta S_h')$ of the tetrahedron (ΔS_h) on the hyperplane with the coordinates $t_{m_1}, t_{m_2}, \ldots, t_{m_p}$ taken with the sign $+$ or $-$ depending on whether the orientation of the projection $\Delta S_h'$ (which is also a tetrahedron) coincides with the orientation of the

tetrahedron $OC_{m_1}C_{m_2}\ldots C_{m_p}$ where C_j $(j = 1, \ldots, k)$ is a point lying on the t_j-axis whose distance from the origin of coordinates is equal to unity. All the indices $m_1, m_2, \ldots, m_p$ are supposed to be different, if otherwise integral (53) is considered, by definition, to be equal to zero.

The properties of integral (53) are analogous to the properties of the integrals investigated in Sec. 3 with the exception of those related to the integration of inequalities. If the orientation of (S) is changed or if any two differentials under the integral sign are permuted the integral is multiplied by -1 (why?). We also consider the sums of integrals of the form

$$\int \ldots \int_{(S)} \sum_{m_1, \ldots, m_p=1}^{k} u_{m_1, \ldots, m_p} (t_1, \ldots, t_k)\, dt_{m_1}\, dt_{m_2} \ldots dt_{m_p} \quad (54)$$

An integral taken over an ordinary two-dimensional oriented surface lying in the ordinary three-dimensional space with the coordinates x, y, z which can be written in the form

$$\int\int_{(S)} P(x, y, z)\, dx\, dy + Q(x, y, z)\, dy\, dz + R(x, y, z)\, dx\, dz$$

is an example of integral (54).

When setting up the limits of integration in integral (53) we can express all the coordinates t_j different from $t_{m_1}, t_{m_2}, \ldots, t_{m_p}$ as functions of $t_{m_1}, t_{m_2}, t_{m_p}$ for the points of the manifold (S). Then we substitute these expressions into the integrand which thus becomes a function of $t_{m_1}, t_{m_2}, \ldots, t_{m_p}$, divide (S) into parts whose projections on the hyperplane with the coordinates $t_{m_1}, t_{m_2}, \ldots, t_{m_p}$ are of the same orientation and then set up the limits of integration in the integrals over each projection. The latter integrals are evaluated as ordinary p-fold multiple integrals taken over a p-dimensional domain. We can also introduce some convenient curvilinear coordinates $s_1, \ldots, s_p$ on (S) and then pass from the differentials $dt_{m_1}, \ldots, dt_{m_p}$ to the differentials $ds_1, \ldots, ds_p$. To do this we substitute the expression

$$\frac{D(t_{m_1}, \ldots, t_{m_p})}{D(s_1, \ldots, s_p)}\, ds_1 \ldots ds_p$$

for $dt_{m_1} \ldots dt_{m_p}$ under the sign of integration and set up the corresponding limits of integration.

§ 6. Vector Field

Multiple integrals are directly applied to the theory of vector field. Here we shall discuss some of the applications. The reader should recall the definition of a field given in Sec. IX.9 before proceeding to study the subject.

21. **Vector Lines.** We say that there is a **vector field** A (field of vector A) defined in space if the value of the vector quantity A is specified at each point M of space, i.e. $\mathbf{A} = \mathbf{A}\,(M)$. We shall deal with a stationary field which does not change as time passes. If such a variation takes place we shall consider the field at a fixed moment of time and thus reduce our considerations to a stationary field. As examples of vector fields, we can consider the field of velocity v, the field of momentum density ρv (where ρ is the density of mass distribution) for a flow of a liquid or gas, the field of force F, the electric field E (where E is the electric field strength) etc.

A curve (L) which is tangent to the vector A at each point is called a **vector line.** In other words, this is a curve whose direction (i.e. the direction of its tangent) coincides with the direction of the field at each point belonging to the curve. Depending on the physical meaning of the field in question we speak about a *stream line* (*flow line*) of a field of velocity, a *line of force* of a field of force and so on. (Let the reader think why the stream lines coincide with the trajectories of the particles of liquid only in the case of a stationary field.)

From the geometrical point of view the problem of constructing vector lines of a given vector field is equivalent to that of constructing integral curves for a given direction field (see Sec. XV.12). The problem is therefore reduced to integrating the corresponding system of differential equations. For this purpose it is necessary to introduce a coordinate system in space. For instance, if we take Cartesian coordinates x, y, z the vector A can be resolved according to the formula

$$\mathbf{A} = \mathbf{A}\,(x,\ y,\ z) = A_x\,(x,\ y,\ z)\,\mathbf{i} + A_y\,(x,\ y,\ z)\,\mathbf{j} + {}$$
$$+ A_z\,(x,\ y,\ z)\,\mathbf{k} \qquad (55)$$

On the basis of Sec. XV.12, we can put down the symmetric system of differential equations for the vector lines of the field A

$$\frac{dx}{A_x\,(x,\ y,\ z)} = \frac{dy}{A_y\,(x,\ y,\ z)} = \frac{dz}{A_z\,(x,\ y,\ z)}$$

[compare this with equation (XV.66)]. In the case of a plane field (see Sec. IX.9) the system turns into the equation

$$\frac{dx}{A_x\,(x,\ y)} = \frac{dy}{A_y\,(x,\ y)}$$

As it was shown in Chapter XV where the theory of differential equations was studied, there is only one vector line passing through a non-singular point. Thus, the whole region in which a vector field is defined is filled with vector lines of the field. In a sufficiently small domain containing a non-singular point the totality of the vector lines resembles the set of parallel segments which can be curved a little. In the vicinity of a singular point the family of vector lines can have a very complicated structure (see Fig. 290 which represents some examples of this kind).
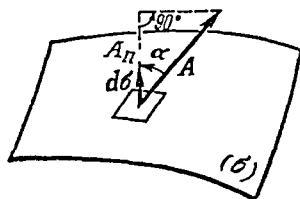
Fig. 323

22. The Flux of a Vector Through a Surface. Let a vector field be defined in a domain of space and let an oriented surface (σ) (which can be closed or non-closed) lie in the domain. We remind the reader that orienting a surface is equivalent to indicating its outer and inner sides (see Sec. VII.11). The flux of a vector field A through a surface (σ) is the surface integral

$$Q = \int_{(\sigma)} A_n \, d\sigma$$

where A_n is the projection of the vector A on the unit outer normal n to (σ). Using the notion of the vector of an area (see Sec. VII.11) and the properties of a scalar product of vectors (see Sec. VII.2) we can rewrite the expression of the flux in the form

$$Q = \int_{(\sigma)} A \cos \alpha \, d\sigma = \int_{(\sigma)} \mathbf{A} \cdot d\boldsymbol{\sigma}$$

(see Fig. 323).

If the vector field A is represented in form (55) we can apply the transformation

$$\mathbf{A} \cdot d\boldsymbol{\sigma} = \mathbf{A} \cdot \mathbf{n} \, d\sigma = (A_x \mathbf{i} + A_y \mathbf{j} + A_z \mathbf{k}) \cdot (\cos(\widehat{\mathbf{n}, x}) \mathbf{i} + \cos(\widehat{\mathbf{n}, y}) \mathbf{j} +$$

$$+ \cos(\widehat{\mathbf{n}, z}) \mathbf{k}) \, d\sigma = A_x \cos(\widehat{\mathbf{n}, x}) \, d\sigma + A_y \cos(\widehat{\mathbf{n}, y}) \, d\sigma +$$

$$+ A_z \cos(\widehat{\mathbf{n}, z}) \, d\sigma$$

to calculating the flux. Thus, the integral $\int_{(\sigma)} \mathbf{A} \cdot d\boldsymbol{\sigma}$ is represented as the sum of three integrals. The first integral $\int_{(\sigma)} A_x \cos(\widehat{\mathbf{n}, x}) \, d\sigma$ can be computed if we use the relation $\cos(\widehat{\mathbf{n}, x}) \, d\sigma = \pm \, d\sigma_x$ where $d\sigma_x$ entering into the right-hand side is the surface element of the projection (σ_x) of the surface (σ) on the y, z-plane and the sign is

40*

buted in space (recall the general concept of a quantity distributed in space which was discussed in § 2). We first turn to the latter case. Here we can introduce not only the average density $\frac{Q}{\Omega}$ of the source in (Ω) [as before, the symbol Ω designates the volume of the domain (Ω)] but also the density of the sources of the field at any point M of space which is defined as

$$\lim_{(\Delta\Omega)\to M} \frac{\Delta Q}{\Delta\Omega} = \lim_{(\Delta\Omega)\to M} \frac{\int\limits_{(\Delta\sigma)} \mathbf{A}\cdot d\sigma}{\Delta\Omega} \tag{56}$$

where $(\Delta\Omega)$ is a small volume enveloping the point M and $(\Delta\sigma)$ is the surface which bounds $(\Delta\Omega)$.

This density is called the **divergence** of the vector field $\mathbf{A}$ and is designated as div $\mathbf{A}$. Hence, we can say that the divergence of a vector field can be interpreted as the number of vector lines generated in the interior of an infinitesimal volume (i.e. the flux of the field through the surface bounding the volume) related to unit volume. It should be noted that the divergence of a vector field is a scalar quantity. Moreover, it forms a scalar field because it assumes a certain numerical value at each point in space.

Formula (56) can be rewritten in the form

$$\operatorname{div}\mathbf{A} = \frac{dQ}{d\Omega}, \quad \text{i.e.} \quad dQ = \operatorname{div}\mathbf{A}\, d\Omega$$

The last expression represents the number of vector lines issued from the element of volume $(d\Omega)$. Summing together these expressions over a domain (Ω) (see Sec. 4) we arrive at the formula for the number of vector lines coming out of the finite volume (Ω) (that is for the flux of the field $\mathbf{A}$):

$$\oint\limits_{(\sigma)} \mathbf{A}\cdot d\sigma = \int\limits_{(\Omega)} \operatorname{div}\mathbf{A}\, d\Omega \tag{57}$$

where (Ω) is any finite domain and (σ) is its boundary surface. This is Ostrogradsky's formula which plays an important role in the vector field theory. It was discovered by Ostrogradsky (in scalar form) in 1826. The formula holds in all cases when the field $\mathbf{A}$ and its divergence div $\mathbf{A}$ do not approach infinity in (Ω). It is also valid when the divergence approaches infinity in such a way that the integral on the right-hand side of formula (57) is convergent.

The physical meaning of the divergence of a field depends on the nature of the vector field $\mathbf{A}$. For instance, by Sec. 22, for the velocity field $\mathbf{v}$ of a gas flow div $\mathbf{v}$ is equal to the rate of relative expansion of an infinitesimal volume of gas and div $(\rho\mathbf{v})$ is equal to the density of mass sources. If the mass of the gas remains constant in the process of its flow we must have div $(\rho\mathbf{v}) = 0$ (in the general case the mass can receive an increment, positive or negative, resulting

from a chemical or some other reaction in which the mass can change). At the same time we can have div $v > 0$, div $v < 0$ or div $v = 0$ depending on whether the gas expands, contracts or does not change its density in the process of flow. If we take an electric field E its divergence, i.e. div E, is proportional to the density of a charge distributed in space and so on.

If a field has sources distributed over curves or surfaces (its volume density must be discontinuous in such a case) we can speak about the line density or the surface density. In such a case we must add to the right-hand side of formula (57) the corresponding line and surface integrals taken over the curves and the surfaces carrying the sources which lie in the domain (Ω). If there are point charges in (Ω) the corresponding summands should also be added to the right-hand side of (57). If we understand the densities as generalized functions which were discussed in Sec. XIV.7 and Sec. 19 formula (57) will be true in all cases.

If we take a plane vector field A the divergence is defined as

$$(\text{div } A)_M = \lim_{(\Delta\sigma)\to M} \frac{\oint_{(\Delta l)} A_n \, dl}{\Delta\sigma} \tag{58}$$

[formula (58) replaces formula (56) in this case]. Here $(\Delta\sigma)$ is a small plane figure enveloping the point M and (Δl) is the contour bounding the figure. As is known (see Sec. IX.9), a plane field can be interpreted in two ways. If the field is defined only in a given plane then, by definition, the numerator on the right-hand side of formula (58) is considered to be the flux of the vector field A across the curve (Δl). But if the field A is originally defined in space and is regarded as a plane field because it does not depend on one of the Cartesian coordinates (for instance, on z) the numerator is equal to the flux of the vector field A through the lateral surface of a right cylinder [with base $(\Delta\sigma)$ and unit height] whose elements are parallel to the z-axis. In this case the denominator on the right-hand side of formula (58) is equal to the volume of the cylinder (why is it so?).

Ostrogradsky's formula for a plane field has the form

$$\oint_{(l)} A_n \, dl = \int_{(\sigma)} \text{div } A \, d\sigma$$

where (σ) is a finite plane figure and (l) is its contour.

The divergence of a field can sometimes be directly computed on the basis of its definition (56). For example, let us consider a centrally symmetric field in space which is defined by the formula

$$A = f(r) \, r^\circ = \frac{f(r)}{r} \, r$$

where **r** is the radius-vector of a variable point and $f(r)$ is a given function of the modulus of r (see Fig. 324). Then the flux across the sphere of radius r is equal to

$$Q(r) = \int A_n \, d\sigma = \int A_r \, d\sigma = \int f(r) \, d\sigma = f(r) \, 4\pi r^2$$

and therefore the number of vector lines issued from a thin spherical layer of width dr is equal to

$$dQ = 4\pi d \, [r^2 f(r)] = 4\pi \, [2rf(r) + r^2 f'(r)] \, dr$$

Consequently, we have

$$\text{div } \mathbf{A} = \frac{2}{r} f(r) + f'(r)$$

which is obtained from the above expression for dQ after it has been divided by the volume $d\Omega = 4\pi r^2 \, dr$ of the layer.

24. **Expressing Divergence in Cartesian Coordinates.** Let a Cartesian coordinate system x, y, z be given in space. Then a vector field **A** can be represented in form (55). In this case we can deduce a simple formula for computing the divergence div **A**. To do this we take into account the fact that the particular
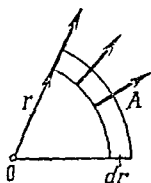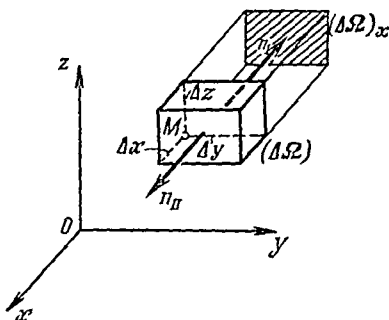
Fig. 324   Fig. 325

form of an infinitesimal domain $(\Delta\Omega)$ entering into the definition of a divergence [see formula (56)] is inessential. Therefore we can take a small rectangular parallelepiped with faces parallel to the coordinate planes as this volume (see Fig. 325). Then the numerator of the fraction in expression (56) can be represented as a sum of six summands corresponding to the six faces of the parallelepiped. We now consider the sum of two summands corresponding to the faces (designated by I and II in Fig. 325) which are parallel to the y, z-plane and whose unit outer normals are denoted as n_I and n_I. We have $(A_n)_I = -(A_x)_I$, and on the basis of Taylor's

formula (see Sec. IV.15) we can write

$$(A_n)_{\text{II}} = (A_x)_{\text{II}} = (A_x)_{\text{I}} + (\partial_x A_x)_{\text{I}} + \cdots$$

where the expression $\partial_x A_x = \dfrac{\partial A_x}{\partial x} \Delta x$ is the partial differential with respect to x which appears here because the points belonging to the faces I and II differ by Δx in the values of their abscissas x. The dots on the right-hand side of the formula designate the terms of higher order of smallness which are not put down here. The integration over these faces reducing to the integration over their projections onto the y, z-plane [i.e. over $(\Delta\Omega)_x$], we have

$$\int_{(\text{I})} A_n \, d\sigma + \int_{(\text{II})} A_n \, d\sigma = \iint_{(\Delta\Omega)_x} (A_n)_{\text{I}} \, dy \, dz + \iint_{(\Delta\Omega)_x} (A_n)_{\text{II}} \, dy \, dz =$$

$$= \iint_{(\Delta\Omega)_x} \left(\frac{\partial A_x}{\partial x}\right)_{\text{I}} \Delta x \, dy \, dz + \cdots = \left(\iint_{(\Delta\Omega)_x} \left(\frac{\partial A_x}{\partial x}\right)_{\text{I}} dy \, dz\right) \Delta x + \cdots =$$

$$= \left(\left(\frac{\partial A_x}{\partial x}\right)_{av} \Delta y \, \Delta z\right) \Delta x + \cdots = \left(\frac{\partial A_x}{\partial x}\right)_{M} \Delta x \, \Delta y \, \Delta z + \cdots$$

The dots here also designate the terms of higher order of smallness relative to the terms which are written in full. The subscripts I, II and M mean that the corresponding terms are taken for the point belonging to the face I, II or for the point M, respectively. The inscription av means the average (mean) value. When performing the last transformation in the above formula we have used the formula

$$\left(\frac{\partial A_x}{\partial x}\right)_{av} = \left(\frac{\partial A_x}{\partial x}\right)_{M} + infinitesimal$$

and in the preceding transformation we have taken advantage of the formula for the mean value of a function (property 10 in Sec. XVI.3).

Performing similar calculations for the other two pairs of faces and summing up all the expressions we arrive at the formula for the flux through the whole boundary surface of the parallelepiped:

$$\int_{(\Delta\sigma)} \mathbf{A} \cdot d\sigma = \left(\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\right)_{M} \Delta x \, \Delta y \, \Delta z + \cdots$$

In this case we have $\Delta\Omega = \Delta x \, \Delta y \, \Delta z$ and therefore

$$\frac{1}{\Delta\Omega} \int_{(\Delta\sigma)} \mathbf{A} \cdot d\sigma = \left(\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\right)_{M} + \cdots$$

Passing to the limit we finally obtain the expression

$$\operatorname{div} \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \tag{59}$$

We have not written the subscript M here because the formula holds for any point of the field.

We can similarly find the expression of a divergence for a plane field. The result must obviously be the same as (59) with the exception that the last term entering into (59) should be deleted.

25. Line Integral and Circulation. Let an oriented curve (L) (i.e. such that the direction of describing this curve is indicated) be given in the domain of space where a vector field **A** is defined. Then we can form the line integral

$$I = \int\limits_{(L)} A_\tau dL \qquad . \qquad (60)$$
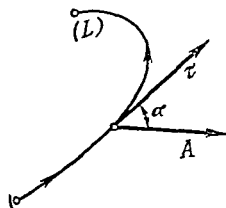
Fig. 326

taken along (L) where A_τ is the projection of the vector **A** on the unit vector τ tangent to (L) drawn in the direction of describing the curve (L) (see Fig. 326). Since the vector dr goes along τ and $|dr| = dL$ (see Sec. VII.23) the expression for line integral (60) can be rewritten in the form

$$I = \int\limits_{(L)} A \cos\alpha \,|\, dr| = \int\limits_{(L)} \mathbf{A} \cdot dr = \int\limits_{(L)} (A_x \, dx + A_y \, dy + A_z \, dz) \quad (61)$$

Line integral (60), or (61), is a scalar quantity possessing all the properties of line integrals (see Sec. XIV.6). If the orientation of the curve (L) is changed the integral is multiplied by -1. If the angle α shown in Fig. 326 is acute at all the points of the curve (L) we have $I > 0$ and if the angle is obtuse $I < 0$. Finally, $I = 0$ if α is a right angle at all points. We can also have $I = 0$ when the angle α is acute for one portion of the curve (L) and is obtuse for the other but varies in such a way that the integrals taken over these portions mutually cancel out.

A line integral of type (60) has an obvious physical meaning when **A** is a field of force. As we showed in Sec. XIV.22, in this case the integral is equal to the work performed by the field when the point upon which the force acts describes the curve (L).

If (L) is a closed curve line integral (60) is called the **circulation** (in this case we can write the integral as $\oint\limits_{(L)} (A_x \, dx + A_y \, dy + A_z \, dz)$.

26. Rotation. For our further aims we need the expression of a circulation taken along an infinitesimal closed contour (ΔL). To obtain the expression we assume that the vector field **A** is represented in form (55), i.e. as being resolved into components along the unit vectors of the Cartesian axes. Let the contour (ΔL) be placed near a point M_0 of space. Now we compute the integral of the first

summand entering into the right-hand side of formula (61):

$$\oint_{(\Delta L)} A_x(x, y, z)\,dx = \oint_{(\Delta L)} \left[(A_x)_0 + \left(\frac{\partial A_x}{\partial x}\right)_0 (x - x_0) + \right.$$

$$\left. + \left(\frac{\partial A_x}{\partial y}\right)_0 (y - y_0) + \left(\frac{\partial A_x}{\partial z}\right)_0 (z - z_0) + \dots \right] dx \qquad (62)$$

Here we have applied Taylor's formula (see Sec. XII.6). The subscript "zero" indicates that the corresponding quantities are taken at the point M_0, and the dots designate the terms of higher order of smallness. Recall that in Sec. XIV.23 we proved the formulas which can be written in the notation of this section as

$$\oint_{(\Delta L)} [C_1 + C_2 x]\,dx = 0,$$

$$\oint_{(\Delta L)} y\,dx = -\Delta S \cos(\widehat{n, z}) \quad \text{and} \quad \oint_{(\Delta L)} z\,dx = \Delta S \cos(\widehat{n, y})$$

where C_1 and C_2 are arbitrary constants, ΔS is the area of the surface (ΔS) bounded by the curve (ΔL) (this surface can be regarded as a plane figure to within infinitesimals of higher order) and n is the unit outer normal to (ΔS) whose direction is coherent with the direction of describing (ΔL) according to the right-hand screw rule (see Sec. VII.11). Hence we can write the relation of the form

$$\oint_{(\Delta L)} \left[(A_x)_0 + \left(\frac{\partial A_x}{\partial x}\right)_0 (x - x_0) - \left(\frac{\partial A_x}{\partial y}\right)_0 y_0 - \left(\frac{\partial A_x}{\partial z}\right)_0 z_0 \right] dx =$$

$$= \oint_{(\Delta L)} [C_1 + C_2 x]\,dx = 0$$

Substituting all these results into (62) we obtain

$$\oint_{(\Delta L)} A_x\,dx = \left[\left(\frac{\partial A_x}{\partial z}\right)_0 \cos(\widehat{n, y}) - \left(\frac{\partial A_x}{\partial y}\right)_0 \cos(\widehat{n, z}) \right] \Delta S + \dots \qquad (63)$$

Evaluating the other two integrals on the right-hand side of formula (61) in a similar manner [to perform the evaluation it is sufficient to make two successive circular permutations of the coordinates in formula (63)] and summing up the results we deduce

$$\oint_{(\Delta L)} \mathbf{A} \cdot d\mathbf{r} = \left\{ \left[\left(\frac{\partial A_x}{\partial z}\right)_0 \cos(\widehat{n, y}) - \left(\frac{\partial A_x}{\partial y}\right)_0 \cos(\widehat{n, z}) \right] + \right.$$

$$\left. + \left[\left(\frac{\partial A_y}{\partial x}\right)_0 \cos(\widehat{n, z}) - \left(\frac{\partial A_y}{\partial z}\right)_0 \cos(\widehat{n, x}) \right] + \right.$$

The rotation of a plane field has a particularly simple expression. Indeed, if $A = A_x(x, y)\,i + A_y(x, y)\,j$ formula (65) implies that in this case we have

$$\operatorname{rot} A = \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right) k$$

27. Green's Formula. Stokes' Formula. These formulas enable us to transform the circulation of a vector field over a closed contour

Fig. 328

to a double integral over a surface bounded by the contour. Green's formula is related to a plane field and Stokes' formula deals with the general case of a spatial field. The former formula directly follows from the latter but we shall deduce Green's formula independently because the deduction is very simple.
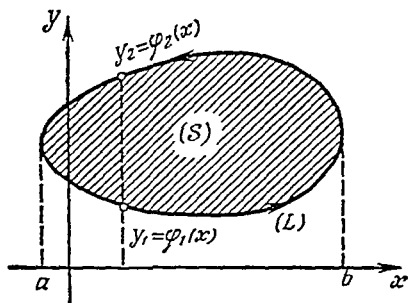
Let us consider the circulation of a plane field $A = P(x, y)\,i + Q(x, y)\,j$ over a closed contour (L) which is described in the positive direction. The finite domain bounded by the contour (L) will be designated by (S) (see Fig. 328). By formula (61), the circulation can be written as

$$\Gamma = \oint_{(L)} P(x, y)\,dx + \oint_{(L)} Q(x, y)\,dy \tag{68}$$

For the first integral in (68) we obtain

$$\int_a^b P(x, y_1)\,dx + \int_b^a P(x, y_2)\,dx = - \int_a^b [P(x, y_2) - P(x, y_1)]\,dx \tag{69}$$

(see Fig. 328). The expression under the sign of integration is a partial increment of P with respect to y which can be represented in the form of an integral of the derivative:

$$P(x, y_2) - P(x, y_1) = \int_{\varphi_1(x)}^{\varphi_2(x)} \frac{\partial P}{\partial y}\,dy$$

Substituting this expression into (69) we obtain

$$- \int_a^b \left(\int_{\varphi_1(x)}^{\varphi_2(x)} \frac{\partial P}{\partial y}\,dy \right) dx = - \iint_{(S)} \frac{\partial P}{\partial y}\,dx\,dy$$

The second integral in (68) is transformed similarly (we leave the calculations to the reader). Adding together the results we arrive at Green's formula:

$$\int\limits_{(L)} (P\,dx + Q\,dy) = \int\limits_{(S)} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dS \qquad (70)$$

The formula is applicable if all the functions P, Q, $\frac{\partial Q}{\partial x}$ and $\frac{\partial P}{\partial y}$ are finite everywhere in (S). In particular, formula (70) implies an assertion mentioned in Sec. XV.6 which states that if the condition $\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$ holds in a simply-connected domain (G) in plane the expression $P\,dx + Q\,dy$ is a total differential in the domain. Actually, by formula (70), we have $\oint\limits_{(L)} (P\,dx + Q\,dy) = 0$ for any closed contour (L) lying in (G) and hence the assertion follows from Sec. XIV.24. The condition that the domain in question should be simply-connected implies that for any contour (L) belonging to (G) the portion of the plane lying inside (L) also belongs to (G) (which may not hold for a multiply-connected domain), the implication being applied to the deduction of the above assertion.
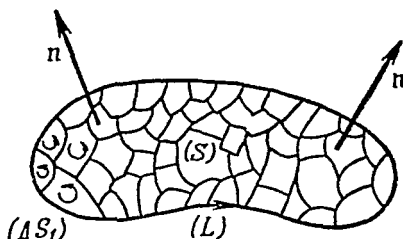
We now proceed to deduce an analogous formula (Stokes' formula)

Fig. 329

for a spatial field.* The formula discovered in 1854 by the English physicist and mathematician G. G. Stokes (1819-1903) is widely applied to the theory of vector field. Let a finite oriented contour (L) bounding a finite oriented surface (S) be given. Let the orientations of (L) and (S) be coherent as it is shown in Fig. 329. Divide (S) into small surfaces (ΔS_1), (ΔS_2), . . ., (ΔS_m) bounded by the contours (ΔL_1), (ΔL_2), . . ., (ΔL_m). The contours (ΔL_i) $(i = 1, \ldots, m)$ are considered to be oriented according to the orientations of (L) and (S). Then we readily conclude that

$$\oint\limits_{(L)} \mathbf{A} \cdot d\mathbf{r} = \sum_{i=1}^{m} \oint\limits_{(\Delta L_i)} \mathbf{A} \cdot d\mathbf{r} \qquad (71)$$

because the integrals taken over the arcs entirely lying inside (L) and which enter into the right-hand side of (71) mutually cancel out (why?) and the sum of the remaining integrals just equals the

* Can be omitted for the first reading of the book.—*Tr.*

left-hand side of formula (71). Regarding (ΔS_i) $(i = 1, \ldots, m)$ as being infinitesimal we can apply formula (66) to each integral $\oint\limits_{(\Delta L_i)} \mathbf{A} \cdot d\mathbf{r}$ which results in

$$\oint\limits_{(L)} \mathbf{A} \cdot d\mathbf{r} = \sum_{i=1}^{m} (\text{rot}_n \mathbf{A})_i \, \Delta S_i + \ldots$$

where the subscript i indicates that the corresponding value of $\text{rot}_n \mathbf{A}$ is taken at a point belonging to the ith area. The sum on the right-hand side is an integral sum (see Sec. 2), and therefore passing to the limit in the process when the linear sizes of all the subdomains are decreased unlimitedly we obtain

$$\oint\limits_{(L)} \mathbf{A} \cdot d\mathbf{r} = \int\limits_{(S)} \text{rot}_n \mathbf{A} \, dS = \int\limits_{(S)} \text{rot} \mathbf{A} \cdot d\mathbf{S} \tag{72}$$

Thus, the circulation of a vector field over a closed contour is equal to the flux of the rotation of the field through a surface bounded by the contour. It is formula (72) that is called **Stokes' formula**.
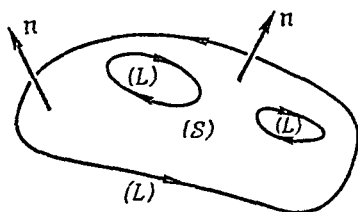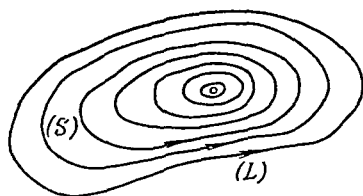
Fig. 330   Fig. 331

The formula holds if the field $\mathbf{A}$ and its rotation are finite on the surface (S). It also holds if the rotation approaches infinity in such a way that the integral on the right-hand side of (72) converges.

We note that a contour (L) entering into Stokes' formula can consist of several portions (components). In this case the contours must be oriented in a corresponding way (see Fig. 330). An analogous remark relates to Ostrogradsky's formula (Sec. 23).

In particular, Stokes' formula implies the sufficiency of conditions (XIV.99) (see Sec. XIV.24 where the question was discussed) for integral (XIV.93) to be independent of the path of integration. Actually, let us introduce the field $\mathbf{A} = P\mathbf{i} + Q\mathbf{j} + R\mathbf{k}$ for which we have $\text{rot} \mathbf{A} = 0$ if conditions (XIV.99) hold. Now taking an arbitrary closed contour (L) and an arbitrary surface (S) spanned by the contour we deduce, on the basis of Stokes' formula, equality (XIV.95). The domain (G) of space in which the whole construction

is performed is supposed to be simply-connected which guarantees the existence of a surface spanned by the contour because if we contract the contour (L) within the above domain (G) in space to a point it describes a surface (S) of the desired type (see Fig. 331).

28. **Expressing Differential Operations on Vector Fields in a Curvilinear Orthogonal Coordinate System.** We now consider a curvilinear orthogonal coordinate system λ, μ, ν in space. It is natural to construct a system of unit vectors e_λ, e_μ, e_ν tangent to the coordinate curves at each point of space and to resolve the vector fields in question with respect to these vectors. Thus we obtain a resolution of the form

$$A = A_\lambda e_\lambda + A_\mu e_\mu + A_\nu e_\nu$$

at each point.

To express the gradient of a scalar field u at an arbitrary point M we should recall that when evaluating the gradient according to formula (XII.2) we can place the Cartesian coordinate system in any way (see Sec. XII.1). Thus, we can put $i = e_\lambda$, $j = e_\mu$ and $k = e_\nu$. This yields

$$\operatorname{grad} u = \frac{\partial_\lambda u}{\partial s_\lambda} e_\lambda + \frac{\partial_\mu u}{\partial s_\mu} e_\mu + \frac{\partial_\nu u}{\partial s_\nu} e_\nu = \frac{1}{l_\lambda} \frac{\partial u}{\partial \lambda} e_\lambda + \frac{1}{l_\mu} \frac{\partial u}{\partial \mu} e_\mu + \frac{1}{l_\nu} \frac{\partial u}{\partial \nu} e_\nu$$

where l_λ, l_μ and l_ν are **Lamé's coefficients** (see Sec. 15).

When calculating the divergence of a vector field we cannot directly apply formula (59) to a curvilinear coordinate system. For instance, if we put $i = e_\lambda$, etc. as above, the equality $A_x = A_\lambda$ will hold only at the point M (why?) and hence the derivative $\frac{\partial A_x}{\partial x}$ cannot be found in such a simple way as above in the general case. Here we can apply the method which was used at the beginning of Sec. 24 in investigating the divergence. Let us consider the flux of the field through the surface of an infinitesimal rectangular parallelepiped bounded by the coordinate surfaces (see Fig. 332). Taking the sum of fluxes across the two faces perpendicular to the coordinate curve λ (on which λ varies whereas μ and ν are constant) we obtain, to within infinitesimals of higher order, the expression

$$\partial_\lambda (A_\lambda \, ds_\mu \, ds_\nu) = \partial_\lambda (l_\mu l_\nu A_\lambda) \, d\mu \, d\nu = \frac{\partial (l_\mu l_\nu A_\lambda)}{\partial \lambda} \, d\lambda \, d\mu \, d\nu$$

Computing the fluxes through the other two pairs of faces, adding together all the results and dividing by the element of volume $d\Omega = l_\lambda l_\mu l_\nu \, d\lambda \, d\mu \, d\nu$ we obtain

$$\operatorname{div} A = \frac{1}{l_\lambda l_\mu l_\nu} \left[\frac{\partial (l_\mu l_\nu A_\lambda)}{\partial \lambda} + \frac{\partial (l_\lambda l_\nu A_\mu)}{\partial \mu} + \frac{\partial (l_\lambda l_\mu A_\nu)}{\partial \nu} \right]$$

To derive the expression for the rotation we must take advantage of formula (67). The circulation of the vector A over the contour of

an infinitesimal rectangle perpendicular to the vector e_λ (see Fig. 333) is equal to

$$\oint A \cdot dr = \left(\int_{NP} - \int_{MQ} \right) - \left(\int_{QP} - \int_{MN} \right) = \partial_\mu (A_\nu \, ds_\nu) - \partial_\nu (A_\mu \, ds_\mu) =$$

$$= \partial_\mu (l_\nu A_\nu \, dv) - \partial_\nu (l_\mu A_\mu \, d\mu) = \left[\frac{\partial (l_\nu A_\nu)}{\partial \mu} - \frac{\partial (l_\mu A_\mu)}{\partial \nu} \right] d\mu \, dv$$

to within infinitesimals of higher order of smallness. Dividing by the surface element $dS = l_\mu l_\nu \, d\mu \, dv$ and performing circular per-
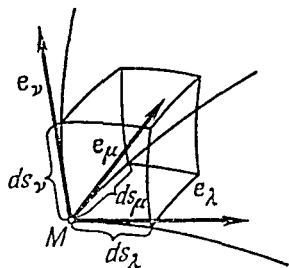
Fig. 332
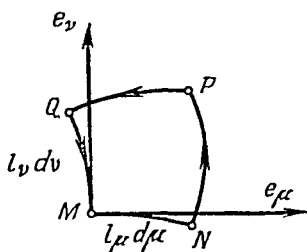
Fig. 333

mutation of the indices we obtain the formulas

$$(\text{rot } A)_\lambda = \frac{1}{l_\mu l_\nu} \left[\frac{\partial (l_\nu A_\nu)}{\partial \mu} - \frac{\partial (l_\mu A_\mu)}{\partial \nu} \right],$$

$$(\text{rot } A)_\mu = \frac{1}{l_\lambda l_\nu} \left[\frac{\partial (l_\lambda A_\lambda)}{\partial \nu} - \frac{\partial (l_\nu A_\nu)}{\partial \lambda} \right],$$

$$(\text{rot } A)_\nu = \frac{1}{l_\lambda l_\mu} \left[\frac{\partial (l_\mu A_\mu)}{\partial \lambda} - \frac{\partial (l_\lambda A_\lambda)}{\partial \mu} \right]$$

All these formulas are naturally simplified in the case of a plane field for which we must put $A_\nu = 0$ and $l_\nu = 1$ and regard all the quantities involved as being independent of ν.

29. General Formula for Transforming Integrals. It turns out that the formulas of Stokes, Ostrogradsky and some formulas in a multiple-dimensional space analogous to them can be written in the form of a single formula which generalizes them all. Let us take the k-dimensional space E_k in which the Lebesgue measure is introduced (see Sec. 20). Consider an oriented $(p + 1)$-dimensional $(p = 1, 2, \ldots, k-1)$ manifold (Ω) with the p-dimensional boundary (Ω') lying in E_k. The orientation of (Ω) induces the corresponding orientation of (Ω') according to the following rule: if a small $(p + 1)$-dimensional tetrahedron $A_1 A_2 A_3 \ldots A_{p+1} A_{p+2}$ which belongs to (Ω) and whose vertices are enumerated in accord with the orientation of (Ω) is placed so that its face $A_1 A_2 A_3 \ldots$

... A_{p+1} belongs to (Ω') this order of the vertices must correspond to the orientation of (Ω'). [Let the reader check that if (Ω) is a surface in the three-dimensional geometric space the above rule coincides with the ordinary rule of the coherence between the orientation of a surface and its contour.]

We now consider integrals of form (54) where $(S) = (\Omega')$. The element of integration

$$\omega = \sum_{m_1, m_2, \ldots, m_p=1}^{k} u_{m_1, \ldots, m_p} (t_1, \ldots, t_k)\, dt_{m_1} \ldots dt_{m_p} \quad (73)$$

is a homogeneous function of degree p with respect to the differentials $dt_1, \ldots, dt_k$. It is called the *differential form of degree p* (*p-form*).

There are some operations which can be performed on differential forms. For instance, we can add together forms of the same degree. By the way, expression (73) is a sum of the simplest forms which are monomials in $dt_1, \ldots$. Differential forms can be multiplied by one another under the convention that, according to the definition of an integral of form (53), a permutation of two differentials entering into a monomial results in multiplying it by -1, and that if there are two similar differentials the monomial is considered, by definition, to be equal to zero. A differential form can also be multiplied by a constant or by a function of $t_1, t_2, \ldots t_k$. By the way, the latter can be regarded as a form of degree zero. The ordinary rules of addition and multiplication hold in this case with the exception that the multiplication of forms is non-commutative in the general case.

A differential form is differentiated according to the following rule:

$$d\omega = d\left(\sum u_{m_1, m_2, \ldots, m_p}\, dt_{m_1} \ldots dt_{m_p} \right) =$$

$$= \sum du_{m_1, \ldots, m_p}\, dt_{m_1} \ldots dt_{m_p} =$$

$$= \sum \left(\frac{\partial u_{m_1, \ldots, m_p}}{\partial t_1}\, dt_1 + \ldots + \frac{\partial u_{m_1, \ldots, m_p}}{\partial t_k}\, dt_k \right) dt_{m_1} \ldots dt_{m_p}$$

where we must remove the brackets and combine the similar terms. We see that the differentiation of a differential form increases its degree by unity.

It turns out that the above definitions imply the general formula for transforming integrals (54):

$$\overbrace{\int \ldots \int}^{p \text{ times}}_{(\Omega')} \omega = \overbrace{\int \int \ldots \int}^{(p+1) \text{ times}}_{(\Omega)} d\omega \quad (74)$$

We shall not give the proof of the formula here.

As an example, let us take the case $k = 2$, $p = 1$. If we write x and y in place of t_1 and t_2 and introduce the notation

$$\omega = P\,(x,\ y)\,dx + Q\,(x,\ y)\,dy$$

we obtain

$$d\omega = dP\,dx + dQ\,dy = \left(\frac{\partial P}{\partial x}\,dx + \frac{\partial P}{\partial y}\,dy\right)dx +$$

$$+ \left(\frac{\partial Q}{\partial x}\,dx + \frac{\partial Q}{\partial y}\,dy\right)dy = \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)dx\,dy$$

It follows that formula (74) turns into a formula which differs from Green's formula (70) only in the notation of the domains of integration. Let the reader consider the case $k = 3$, $p = 1$ (which leads to Stokes' formula) and the case $k = 3$, $p = 2$ (which yields Ostrogradsky's formula). In treating these cases the reader should take into account the following expression of a flux in the form of a double integral with respect to coordinates:

$$\int\limits_{(\sigma)} \mathbf{A}\cdot d\mathbf{\sigma} = \int\limits_{(\sigma)} \mathbf{A}\cdot\mathbf{n}\,d\sigma = \int\limits_{(\sigma)} [A_x\cos(\widehat{n,\ x}) + A_y\cos(\widehat{n,\ y}) +$$

$$+ A_z\cos(\widehat{n,\ z})]\,d\sigma = \int\limits_{(\sigma)}\int (A_x\,dy\,dz + A_y\,dz\,dx + A_z\,dx\,dy)$$

(compare this expression with formulas given in Sec. 22).

Series

We have already dealt with series in our course. We suggest that the reader should look through Sec. III.6, where the basic definitions of the convergence and of the sum of an infinite series were formulated, before proceeding to study the present chapter. Here we shall give a systematic representation of the theory of series.

§ 1. Number Series

1. Positive Series. We now consider a series of the form

$$a_1 + a_2 + \ldots + a_n + \ldots \qquad (1)$$

in which $a_n \geqslant 0$ for all $n = 1, 2, 3, \ldots$. Such a series with non-negative terms is referred to as **positive series**. As in Sec. III.6, let us denote the partial sums of the series as $S_1, S_2, \ldots, S_n, \ldots$. In this case we have $S_1 \leqslant S_2 \leqslant \ldots \leqslant S_n \leqslant \ldots$ (why?). Therefore, recalling the two possible ways of variation of an increasing quantity (see Sec. III.5) we conclude that series (1) is either convergent or **properly divergent** (i.e. divergent to infinity), its sum being $+\infty$ in the latter case. This can be written as

$$\sum_{h=1}^{\infty} a_h < \infty \quad \text{or} \quad \sum_{h=1}^{\infty} a_h = \infty$$

respectively.

It should be noted that the first inequality is a symbolical expression of the convergence of the series and it makes sense only for positive series.

If besides series (1) we consider a series

$$b_1 + b_2 + \ldots + b_n + \ldots \qquad (2)$$

such that

$$0 \leqslant a_h \leqslant b_h \quad (k = 1, 2, 3, \ldots) \qquad (3)$$

then

$$\sum_{k=1}^{\infty} a_k \leqslant \sum_{k=1}^{\infty} b_k$$

Indeed, this is directly implied by the analogous inequality for the partial sums of the series. Thus, we arrive at the comparison test for the convergence of a series similar to the one given in Sec. XIV.15 for an improper integral: if condition (3) holds the convergence of series (2) implies the convergence of series (1) and the divergence of series (1) implies the divergence of series (2).

For example, the series

$$\frac{1}{3^2 \ln 2} + \frac{1}{3^3 \ln 3} + \frac{1}{3^4 \ln 4} + \cdots$$

converges which follows from the comparison of this series with series (III.6):

$$\frac{1}{3^n \ln n} < \frac{1}{3^n} \quad (n = 3, 4, \ldots)$$

Although the first terms of the series do not satisfy the above inequality this does not affect the convergence (see Sec. III.6).

The comparison test implies an analogous test: if

$$a_k > 0, \ b_k > 0 \quad (k = 1, 2, \ldots) \quad \text{and}$$

$$\frac{a_k}{b_k} \xrightarrow[k \to \infty]{} C \quad \text{where} \quad C = \text{const} \neq 0 \quad \text{and} \quad C \neq \infty$$

series (1) and (2) converge or, respectively, diverge simultaneously. Actually, the above condition implies that the ratio $\frac{a_k}{b_k}$ lies between some constant positive limits m and M for all k:

$$m \leqslant \frac{a_k}{b_k} \leqslant M, \quad \text{i.e.} \quad m b_k \leqslant a_k \leqslant M b_k$$

Now, summing with respect to k from 1 to n and then passing to the limit for $n \to \infty$, we obtain

$$m \sum_{k=1}^{\infty} b_k \leqslant \sum_{k=1}^{\infty} a_k \leqslant M \sum_{k=1}^{\infty} b_k$$

which implies our assertion (why?).

Now let us formulate D'Alembert's test which is sufficient for the convergence of series (1) and is widely applied: if the limit

$$\lim_{n \to \infty} \frac{a_{n+1}}{a_n} = l$$

exists for series (1), the series converges in the case $l < 1$ and diverges in the case $l > 1$. The latter assertion is obvious. Indeed,

if $l > 1$ the ratio $\frac{a_{n+1}}{a_n}$ (which approaches l as n increases) becomes greater than unity for sufficiently large n. Hence, the terms of the series increase together with n when n becomes sufficiently large and therefore the necessary condition for the convergence of a series (see Sec. III.6) is violated. Now let us turn to the case $l < 1$. Choose a constant number l' between l and 1. The ratio approaching l indefinitely, we have $\frac{a_{n+1}}{a_n} < l'$ for $n \gg N$ where N is a fixed number. Thus, we obtain

$$\frac{a_{N+1}}{a_N} < l', \quad \frac{a_{N+2}}{a_{N+1}} < l', \quad \frac{a_{N+3}}{a_{N+2}} < l', \ldots$$

which implies

$$a_{N+1} < a_N l', \quad a_{N+2} < a_{N+1} l' < a_N l'^2,$$
$$a_{N+3} < a_{N+2} l' < a_N l'^3 \quad \text{etc.}$$

Hence, after some number N, the terms of series (1) are smaller than the corresponding terms of the series

$$a_N + a_N l' + a_N l'^2 + a_N l'^3 + \ldots$$

The quantity l' satisfying the inequality $0 < l' < 1$, the latter series is a geometric series (progression) with common ratio $l' < 1$ which converges [see series (III.7)]. Hence, by the comparison test, series (1) also converges.

Let us consider an example. To apply D'Alembert's test to the series

$$\sum_{n=1}^{\infty} \frac{a^n}{n^p} \quad (a > 0, \quad p > 0 \quad \text{or} \quad p < 0) \tag{4}$$

it is necessary to consider the limit

$$\lim_{n \to \infty} \left(\frac{a^{n+1}}{(n+1)^p} : \frac{a^n}{n^p} \right) = \lim_{n \to \infty} \frac{a}{\left(1 + \frac{1}{n}\right)^p} = a$$

Hence, series (4) converges for $a < 1$ and diverges for $a > 1$. In the case $a = 1$ D'Alembert's test does not enable us to find out whether the series converges or not.

When D'Alembert's test fails to give the answer it is sometimes possible to apply **Cauchy's integral test** which is more general. It gives the following condition sufficient for the convergence of a positive series: if the expression a_n can be defined not only for the integral values of n ($n = 1, 2, 3, \ldots$) but also for all real $n \gg 1$ and if a_n decreases when n increases series (1) and the integral $\int_1^{\infty} a_n \, dn$ converge or diverge simultaneously. To prove the test let

us establish the inequalities

$$\int\limits_{1}^{\infty} a_n \, dn \leqslant \sum\limits_{n=1}^{\infty} a_n \leqslant \int\limits_{1}^{\infty} a_n \, dn + a_1 \qquad (5)$$

which directly imply the above assertion according to the comparison test. To obtain (5) we write, on the basis of Fig. 334a, the inequality
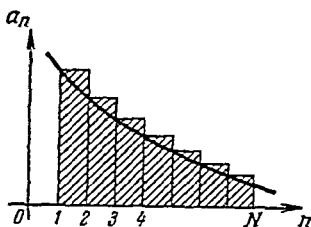
$$\int\limits_{1}^{N} a_n \, dn \leqslant a_1 \cdot 1 + a_2 \cdot 1 + \ldots + a_N \cdot 1 \quad (6)$$

Similarly, Fig. 334b shows that
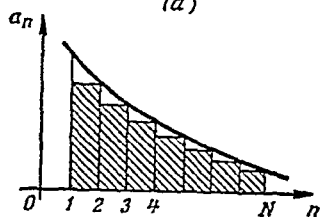
$$\int\limits_{1}^{N} a_n \, dn \geqslant a_2 \cdot 1 + a_3 \cdot 1 + \ldots + a_N \cdot 1$$

and hence

$$a_1 + a_2 + a_3 + \ldots + a_N \leqslant \int\limits_{1}^{N} a_n \, dn + a_1 \qquad (7)$$

(a)

(b)

Fig. 334

If we pass to the limit as $N \to \infty$ in inequalities (6) and (7) we just arrive at (5).

As an example, let us take the series

$$\sum\limits_{n=1}^{\infty} \frac{1}{n^p} \qquad (8)$$

which is a particular case of series (4) for $a = 1$. D'Alembert's test does not give the answer whether the series converges because in this case we have $\lim\limits_{n \to \infty} \frac{a_{n+1}}{a_n} = 1$. But the integral test is applicable here. Indeed, if we consider $a_n = n^{-p}$ as a continuous function of n for all real values of n (exceeding unity) we see that it satisfies the conditions of Cauchy's integral test for $p > 0$ and therefore series (8) and the integral

$$\int\limits_{1}^{\infty} \frac{1}{n^p} \, dn$$

converge or diverge simultaneously.

In Sec. XIV.15 we showed that the last integral converges only when $p > 1$ [see the computation of integral (XIV.51)]. Thus, series (8) converges only in the case $p > 1$. In particular, for $p = 1$ we obtain the so-called **harmonic series**

$$1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} + \ldots = \infty \quad .$$

Formulas (6) and (7) make it possible to estimate a partial sum of a divergent series which enables us to derive an asymptotic formula for such a sum depending on its number. We can similarly estimate any sum of a large number of summands which monotonically depend on the number. To specify the result we can separately sum up a number of the greatest summands in a direct manner and apply the above method of estimation only to the remaining summands because this will decrease the difference between the upper and lower bounds entering into a formula of type (5).

More accurate approximations can be obtained by applying formulas of numerical integration (see Sec. XIV.13). For instance, let us illustrate the application of Simpson's formula to an approximate computation of the sum

$$S_{m,N} = \frac{1}{m} + \frac{1}{m+1} + \ldots + \frac{1}{N} \quad (m = 1, 2, \ldots; \quad N \geqslant m + 2)$$

To do this we write, on the basis of formula (XIV.39) taken for $h = 1$, the relation

$$\int_k^{k+1} \frac{1}{x} \, dx + \int_{k+1}^{k+2} \frac{1}{x} \, dx = \int_k^{k+2} \frac{1}{x} \, dx \approx \frac{1}{3} \left(\frac{1}{k} + \frac{4}{k+1} + \frac{1}{k+2} \right)$$

Adding together these results for $k = m$, $m + 1, \ldots, N - 1$ and performing some simple transformations and the integration we obtain the approximate equality

$$\ln N - \ln m + \ln (N + 1) - \ln (m + 1) \approx$$

$$\approx \frac{1}{3} \left(6 S_{m,N} - \frac{5}{m} + \frac{1}{m+1} - \frac{1}{N} + \frac{1}{N+1} \right)$$

This implies

$$S_{m,N} \approx \ln N + \frac{6m+5}{6m(m+1)} - \frac{1}{2} \ln (m^2 + m) +$$

$$+ \frac{1}{2} \ln \left(1 + \frac{1}{N} \right) + \frac{1}{6N(N+1)} \tag{9}$$

Fig. 334a obviously indicates that for the function $a_n = \frac{1}{n}$ there exists a finite positive limit of the form

$$C = \lim_{N \to \infty} \left[\frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{N} - \int_1^N \frac{1}{n} \, dn \right] = \lim_{N \to \infty} (S_{1,N} - \ln N)$$

which is called **Euler's constant.** Equality (9) and formula $S_{m, N} = = S_{1, N} - \frac{1}{1} - \frac{1}{2} - \ldots - \frac{1}{m-1}$ imply an approximate expression for Euler's constant of the form

$$C \approx \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{m-1} + \frac{6m+5}{6m(m+1)} - \frac{1}{2} \ln(m^2 + m)$$

whose accuracy increases with the growth of m. For $m = 1$ and $m = 2$ we get the approximate values 0.570 and 0.576, respectively, of Euler's constant. The calculations show that $C = 0.5772$ to within 10^{-4}, and thus the above values are accurate to two decimal places.

For greater detail on the problem of computing sums by means of integrals the reader is referred to [51] (see §§ I.2 and III.4).

2. **Series with Terms of Arbitrary Signs.** We now turn to series of the form

$$a_1 + a_2 + \ldots + a_n + \ldots \tag{10}$$

whose term can be of any sign. Let us form the positive series

$$|a_1| + |a_2| + \ldots + |a_n| + \ldots$$

with terms equal to the absolute values of the terms of series (10). We assert that if

$$\sum_{k=1}^{\infty} |a_k| < \infty \tag{11}$$

series (10) is convergent. In this oase series (10) is said to be **absolutely convergent.** The proof is quite similar to that of an analogous property discussed in Sec. XIV.15 and we leave it to the reader. If series $|a_1| + |a_2| + \ldots + |a_n| + \ldots$ diverges series (10) may nevertheless converge. In such a case we say that series (10) is **conditionally convergent.**

When we are given a general series of form (10) we can apply the tests given in Sec. 1 to the series of the moduli of its terms and thus investigate whether it absolutely converges or not. For instance, if

$$\lim_{n \to \infty} \frac{|a_{n+1}|}{|a_n|} < 1$$

the series $|a_1| + |a_2| + \ldots + |a_n| + \ldots$ converges according to D'Alembert's test and hence series (10) converges absolutely. If the limit exceeds unity the necessary test for convergence of a series is violated and thus series (10) diverges.

The following test is referred to as the **Leibniz test.** It is applied to the so-called **alternating series,** i.e. series of the form

$$a_1 - a_2 + a_3 - a_4 + \ldots \tag{12}$$

where $a_k > 0$ $(k = 1, 2, \ldots)$. The Leibniz test asserts that if $a_1 > a_2 > a_3 > \ldots$ and $\lim\limits_{k \to \infty} a_k = 0$ series (12) converges. To prove the test let us mark the points corresponding to the numerical values of the partial sums of series (12) on the S-axis (see Fig. 335). Then, considering the transitions from 0 to S_1, from S_1 to S_2, from S_2 to S_3 etc. we see that every subsequent transition is performed in the direction opposite to that of the preceding transition and at the same time the corresponding distances between the points S_k
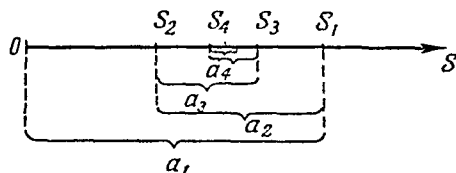
Fig. 335

and S_{k+1} $(k = 1, 2, \ldots)$ decrease. Thus, we have $0 < S_2 < S_1$, $S_2 < S_3 < S_1$, $S_2 < S_4 < S_3$, $\ldots$ (see Fig. 335). Consequently, the partial sums with even numbers form an increasing bounded sequence and therefore they have a finite limit S' (see property 10 in Sec. III.5). Similarly, the partial sums with odd numbers form a decreasing bounded sequence and have a limit S''. Taking the equality $S_{2n+1} = S_{2n} + a_{2n+1}$ and passing to the limit as $n \to \infty$ we obtain $S'' = S'$. Hence, all the partial sums have the same limit and thus series (12) converges. Incidentally, we conclude that the sum of series (12) lies between any sum with an even number and any sum with an odd number which enables us to estimate the sum of the series.

For example, the series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^p} \ .$$

satisfies the conditions of the Leibniz test for $p > 0$ and therefore it converges for such p. For $p > 1$ the convergence will be absolute because, as we know, the series $\sum\limits_{n=1}^{\infty} \frac{1}{n^p}$ converges for $p > 1$ (see Sec. 1). But in the case $p \leqslant 1$ the series $\sum\limits_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^p}$ is conditionally convergent, i.e. it is convergent but not absolutely convergent.

In conclusion, after passing to the limit in the inequality

$$\left|\sum_{h=1}^{n} a_h\right| \leqslant \sum_{h=1}^{n} |a_h|$$

we see that if series (10) converges it satisfies the inequality

$$\left|\sum_{h=1}^{\infty} a_h\right| \leqslant \sum_{h=1}^{\infty} |a_h|$$

3. Operations on Series.

1. Convergent series can be termwise added together (or subtracted), that is if

$$a_1 + a_2 + \ldots + a_n + \ldots = S \quad \text{and}$$
$$b_1 + b_2 + \ldots + b_n + \ldots = T$$

we have

$$(a_1 \pm b_1) + (a_2 \pm b_2) + \ldots + (a_n \pm b_n) + \ldots = S \pm T$$

To prove this we must take the obvious expression $P_n = S_n \pm T_n$ of the sum of the latter series and then pass to the limit as $n \to \infty$.

This property enables us to perform the following transformation of a series. Suppose we are given an absolutely convergent series with terms of arbitrary signs. Let us put it down in the form

$$a - b - c + d + e + f - g + \ldots = S \tag{13}$$

where all a, b, c, $\ldots$ are positive. Let us form the positive series

$$\left.\begin{array}{l} a + 0 + 0 + d + e + f + 0 + \ldots = S_1 \\ 0 + b + c + 0 + 0 + 0 + g + \ldots = S_2 \end{array}\right\} \tag{14}$$

Here we have separately added together all the positive terms and all the absolute values of the negative terms of the original series. Then S is equal to the difference $S_1 - S_2$ because we can termwise subtract the second series from the first one. This operation can be performed only on absolutely convergent series because both series (14) diverge for a conditionally convergent series of form (13) (why?). In the case of a conditionally convergent series the partial sums of both series (14) tend to infinity and the conditional convergence of series (13) is due to the "balance" between these infinities, i.e. the difference between the partial sums tends to zero although the sums themselves are infinitely large.

The next property can be verified in a similar way.

2. A convergent series can be multiplied termwise by a constant factor, i.e. if

$$a_1 + a_2 + \ldots + a_n + \ldots = S$$

we have

$$ka_1 + ka_2 + \ldots + ka_n + \ldots = kS$$

3. We can arbitrarily group the terms when summing a convergent series; for instance, if

$$a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + \ldots = S \qquad (15)$$

we have

$$(a_1 + a_2) + a_3 + (a_4 + a_5 + a_6) + (a_7 + a_8) + \ldots = S \qquad (16)$$

Indeed, if the partial sums S_1, S_2, S_3, ... of the former series tend to S, the partial sums of the latter series which respectively equal S_2, S_3, S_6, S_8, ... also tend to S.

If series (15) properly diverges and we have $S = \infty$ the same is true for series (16). But if (15) is an **oscillating divergent series** (see Sec. III.6) series (16) can diverge or converge and its sum will depend on the way of grouping the terms, i.e. on the way of bracketing. For instance, for series (III.9) we have

$$(1 - 1) + (1 - 1) + (1 - 1) + \ldots = 0 + 0 + 0 + \ldots = 0$$

and

$$1 + (-1 + 1) + (-1 + 1) + \ldots = 1 + 0 + 0 + \ldots = 1$$

Before the difference between convergent and divergent series was understood the above fact had been thought of as an inexplicable paradox. The modern definition of the sum of a convergent series was formulated by Cauchy in 1821, after the theory of limits was created, although series were widely used as early as the 17th and 18th centuries.

4. We can arbitrarily rearrange the terms in a positive series without affecting its sum.

Indeed, if we arbitrarily change the order of terms in a positive series (without omitting any of them) and take successive partial sums of the new series, any term of the original series will enter into all the sums with sufficiently large numbers. Consequently, any partial sum of the original series will be a part of a partial sum of the new series having a sufficiently large number. This implies that the limit of the partial sums of the original series, i.e. its sum, does not exceed the sum of the new series. The original series can be obtained from the new series by rearranging the terms of the latter and therefore the same argument shows that the new sum cannot exceed the old one. Hence, these sums are equal.

An absolutely convergent series with terms of arbitrary signs can also be rearranged in an arbitrary way without affecting the sum.

Actually, as it was shown in property 1, an absolutely convergent series can be represented as a difference of two positive convergent

series. Hence, any rearrangement of the terms of the original series reduces to a rearrangement of the terms of these two series which, as it has been shown, does not affect their sums.

Rearranging the terms of a conditionally convergent series we can make it converge to any sum and even make it diverge. The thing is that the partial sums of the positive and negative terms [see (14)] of a conditionally convergent series are in a balance in the sense that their rates of growth are of the same order. When rearranging the terms of such a series we can change the relation between these rates which can lead to the above result. At first glance this fact looks like a paradox. We shall illustrate what has been said by giving a simple example.

Take the series

$$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\frac{1}{6}+\frac{1}{7}-\frac{1}{8}+\frac{1}{9}-\frac{1}{10}+\frac{1}{11}-\ldots=S \quad (17)$$

According to Sec. 2, its sum lies between the limits $S_2 = 0.5$ and $S_1 = 1$. By property 2, this implies that

$$\frac{1}{2}-\frac{1}{4}+\frac{1}{6}-\frac{1}{8}+\frac{1}{10}-\ldots=\frac{S}{2}$$

and therefore we also have

$$0+\frac{1}{2}+0-\frac{1}{4}+0+\frac{1}{6}+0-\frac{1}{8}+0+\frac{1}{10}+0-\frac{1}{12}+\ldots=\frac{S}{2}$$

Adding termwise the last series and series (17) we obtain

$$1+0+\frac{1}{3}-\frac{1}{2}+\frac{1}{5}+0+\frac{1}{7}-\frac{1}{4}+\frac{1}{9}+0+\frac{1}{11}-\frac{1}{6}+\ldots=\frac{3}{2}S$$

that is

$$1+\frac{1}{3}-\frac{1}{2}+\frac{1}{5}+\frac{1}{7}-\frac{1}{4}+\frac{1}{9}+\frac{1}{11}-\frac{1}{6}+\ldots=\frac{3}{2}S$$

But the above series can be obtained from series (17) by rearranging the terms of the latter (check it up!) and hence we see that the sum has been changed.

Thus, when dealing with processes connected with a rearrangement of the terms of a series we can treat the absolutely convergent series as if they were finite sums. At the same time we must cautiously perform such operations on conditionally convergent series.

4. Speed of Convergence of a Series. In practical computations of the sum of a series we usually compute the sum of several terms suppressing the others when there is no reason to think that these terms can essentially affect the sum (compare with the computation of number e in Sec. IV.16). For this method to yield a good result it is necessary that the series in question not just converge but converge fast so that we could exhaust almost the whole sum when

taking a small number of terms, that is obtain a result approximating the sum with a sufficient accuracy. If a series converges slowly it is usually inapplicable to practical computations. By the way, it is sometimes possible to obtain a new series by performing some operation on the original series or to derive an approximate expression for its remainder by applying integrals as it was done in Sec. 1. Conditionally convergent series usually converge very slowly (see Sec. 2). But there are also absolutely convergent series which converge slowly.

The speed of convergence of a series is essentially dependent on the rate of variation of its general term when it tends to zero, as the number n increases. Series whose general term a_n is of the order of n^{-p} [i.e. $a_n = O(n^{-p})$; see Sec. III.11] for $p > 1$ usually converge slowly. The greater p, the better the convergence of such series. Series with terms a_n of the order of q^n, $0 < q < 1$, converge faster. Their convergence can be compared with that of a geometric series of form (III.7), and the smaller q, the faster the convergence. The convergence of series whose terms a_n are of the order of $\frac{1}{n!}$ is still better etc.

Of course, what has been said represents only general considerations concerning the speed of convergence, and in a concrete case not only the behaviour of the general term for $n \to \infty$ can be essential but also the character of the first terms of the series.

When we are given a slowly convergent series

$$a_1 + a_2 + \ldots + a_n + \ldots \qquad (18)$$

we can try to pass from it to a series which converges faster. One of these methods is as follows. We choose a series

$$b_1 + b_2 + \ldots + b_n + \ldots = \sigma$$

whose sum σ is known so that $a_n \sim b_n$ for $n \to \infty$ (see Secs. III.7, 8). Then we have $a_n = b_n + \gamma_n$ where $|\gamma_n| \ll |a_n|$, and therefore series (18) can be represented in the form

$$(b_1 + \gamma_1) + (b_2 + \gamma_2) + \ldots = (b_1 + b_2 + \ldots) + (\gamma_1 + \gamma_2 + \ldots) =$$
$$= \sigma + \gamma_1 + \gamma_2 + \ldots + \gamma_n + \ldots$$

and the general term of the latter series tends to zero faster than that of the original series.

To apply the above method we must have a set of series with known sums at our disposal. Geometric series (III.7), series indicated in Sec. IV.16, some combinations of these series and the series

$$\sum_{n=1}^{\infty} \frac{1}{n^p} = \zeta(p) \qquad (p > 1) \qquad (19)$$

are most frequently used for this purpose. The latter sum (dependent on p) is called the **Riemann zeta function** after the prominent German mathematician G. F. B. Riemann (1826-1866) although it was Euler who was the first to introduce the function in 1737. The tables of the values of the function can be found in [23].

For instance, let us take the series

$$S = \sum_{n=1}^{\infty} \frac{1}{\sqrt{n^3 + 1}} \tag{20}$$

Its terms are equivalent (as infinitesimals) to the terms of series (19) for $p = \frac{3}{2}$, as $n \to \infty$. Hence, series (20) converges but very slowly. Taking advantage of inequalities (5) we can readily verify that the remainder after n terms of series (19) is equivalent to $\frac{1}{(p-1)\, n^{p-1}}$, i.e. the remainder of series (20) is of the order of $2n^{-\frac{1}{2}}$. Thus to obtain its sum S with an accuracy of 0.01 we must take about 40,000 terms! But here we can apply the above method and write

$$\frac{1}{\sqrt{n^3 + 1}} = \frac{1}{\sqrt{n^3}} + \gamma_n$$

where

$$\gamma_n = \frac{\sqrt{n^3} - \sqrt{n^3 + 1}}{\sqrt{n^3 + 1}\,\sqrt{n^3}} = - \frac{1}{\sqrt{n^3}\,\sqrt{n^3 + 1}\,(\sqrt{n^3} + \sqrt{n^3 + 1})}$$

Consequently, series (20) can be represented in the form

$$S = \zeta\left(\frac{3}{2}\right) - \sum_{n=1}^{\infty} \frac{1}{\sqrt{n^3}\,\sqrt{n^3 + 1}\,(\sqrt{n^3} + \sqrt{n^3 + 1})} \tag{21}$$

The tabular value of the first summand is equal to 2.612, and the general term of the latter series is equivalent to $(2n^{\frac{9}{2}})^{-1}$. Hence, the remainder of the last series is of the order of $(7n^{\frac{7}{2}})^{-1}$ which means that to obtain its sum to within 0.01 it is sufficient to take only three terms! If a greater accuracy is needed we can repeatedly apply the method which yields

$$S = \zeta\left(\frac{3}{2}\right) - \frac{1}{2}\,\zeta\left(\frac{9}{2}\right) +$$

$$+ \sum_{n=1}^{\infty} \frac{24n^9 + 7n^6 - 2n^3 - 1}{2\,\sqrt{n^9}A\,(\sqrt{n^3} + A)\,(3n^3 + 1 + \sqrt{n^3}\,A)\,(2\,\sqrt{n^3}\,A^3 + 2n^6 - 3n^3 - 1)}$$

(check up the result!) where we have put, for brevity, $\sqrt{n^3 + 1} = A$. The sum of the tabular values of the first two terms is equal to

2.085, and the remainder of the last series is equivalent to $\frac{3}{52} n^{-\frac{13}{2}}$. Hence, to obtain S with an accuracy of 0.001 it is sufficient to take only two terms.

The above successive application of the method can be perfected if we apply Taylor's series (IV.60) (the binomial series) to the expression $\frac{1}{\sqrt{n^3+1}}$:

$$\frac{1}{\sqrt{n^3+1}} = (n^3+1)^{-\frac{1}{2}} = n^{-\frac{3}{2}} \left(1+\frac{1}{n^3}\right)^{-\frac{1}{2}} =$$

$$= n^{-\frac{3}{2}} \left(1 - \frac{1}{2n^3} + \frac{3}{8n^6} - \frac{5}{16n^9} + \dots\right). \tag{22}$$

Suppressing the terms of series (22) following after any term we obtain the corresponding approximation. For instance, dropping the terms after the third, we obtain

$$\frac{1}{\sqrt{n^3+1}} = n^{-\frac{3}{2}} \left(1 - \frac{1}{2n^3} + \frac{3}{8n^6}\right) - \gamma_n \tag{23}$$

which yields

$$S = \zeta\left(\frac{3}{2}\right) - \frac{1}{2}\zeta\left(\frac{9}{2}\right) + \frac{3}{8}\zeta\left(\frac{15}{2}\right) - \sum_{n=1}^{\infty} \gamma_n = 2.462 - \sum_{n=1}^{\infty} \gamma_n$$

The terms γ_n can be found from (23). On the basis of (22), we conclude that they are equivalent to $\frac{5}{16} n^{-\frac{21}{2}}$.

Many other series can be transformed in a similar way. Besides (19) we also use the series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^p} = \sum_{n=1}^{\infty} \frac{1}{n^p} - \sum_{n=1}^{\infty} \frac{2}{(2n)^p} = \zeta(p)\left(1 - \frac{1}{2^{p-1}}\right), \tag{24}$$

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1}\right) =$$

$$= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \dots = 1,$$

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)(n+2)} = \frac{1}{2} \sum_{n=1}^{\infty} \left[\frac{1}{n(n+1)} - \frac{1}{(n+1)(n+2)}\right] =$$

$$= \frac{1}{2}\left[\left(\frac{1}{1\cdot 2} - \frac{1}{2\cdot 3}\right) + \left(\frac{1}{2\cdot 3} - \frac{1}{3\cdot 4}\right) + \dots\right] = \frac{1}{4}$$

and so on.

Formula (24) also holds for $0 < p < 1$ but the function $\zeta(p)$ cannot be defined by formula (19) in this case since the series diverges, and it is therefore defined by means of another method which we shall not discuss here. Formula (IV.61) implies that, for $p = 1$, the left-hand side of formula (24) is a convergent series whose sum is equal to $\ln 2$.

When we deal with alternating series it is usually very difficult to guarantee the desired accuracy in practical computations. For example, let us take series (IV.56) for the cosine putting $x = 100$:

$$\cos 100 = 1 - \frac{100^2}{2!} + \frac{100^4}{4!} - \frac{100^6}{6!} + \frac{100^8}{8!} - \ldots \qquad (25)$$

The series on the right-hand side of (25) is convergent and even absolutely convergent (why?). But we cannot use it for practical calculations. The thing is that although its terms beginning with the 51st decrease sufficiently fast so that the theoretical convergence is guaranteed they become enormously large before that. The whole sum does not exceed unity in its absolute value and hence all these large terms must almost completely mutually cancel. As we know from Sec. I.9, in such circumstances, in order to achieve the desired accuracy, we must carry out all the calculations with a great number of significant digits and hence perform much unnecessary work. Therefore we should avoid using series of type (25). If such series occur we must transform them to other series convenient for practical calculations. For instance, in the above example we can take advantage of the periodicity of the cosine and pass to a considerably smaller value of the argument.

5. Series with Complex, Vector and Matrix Terms. The definition of the convergence and of the sum of a series with complex terms

$$z_1 + z_2 + \ldots + z_n + \ldots, \quad z_n = x_n + iy_n \quad (n = 1, 2, 3, \ldots) \qquad (26)$$

is completely the same as that of real series (see Sec. III.6). Series (26) is usually reduced to two series of the form

$$x_1 + x_2 + \ldots + x_n + \ldots \quad \text{and} \quad y_1 + y_2 + \ldots + y_n + \ldots \qquad (27)$$

If both series (27) converge and have the sums x and y, respectively, series (26) also converges and has the sum $z = x + iy$. If at least one of the series (27) is divergent series (26) is divergent as well. Series (27) having real terms, we can apply the methods of Sec. 2 to them.

The following test is also of use: if

$$\sum_{n=1}^{\infty} |z_n| < \infty \qquad (28)$$

both series (27) are absolutely convergent and therefore series (26) is also convergent. If (28) holds series (26) is said to be *absolutely*

convergent. The methods of Sec. 1 can be applied to a series satisfying condition (28).

A series of the form

$$u_1 + u_2 + \ldots + u_n + \ldots \tag{29}$$

whose terms are vectors is treated in like manner. If all the vectors u_n $(n = 1, 2, \ldots)$ belong to a three-dimensional space we can pass to the corresponding series with scalar terms by projecting series (29) on the x, y and z-axes.

We can also consider a series of the form

$$A_1 + A_2 + \ldots + A_n + \ldots \tag{30}$$

where the terms A_n $(n = 1, 2, \ldots)$ are matrices. For series (30) to converge it is necessary and sufficient that each of the series formed of the corresponding elements of the matrices (Sec. XI.2) (that is of the elements standing at the intersections of the same rows and columns of the matrices) should converge.

The properties of series (26), (29) and (30) are the same as those of real series (see Sec. 3).

6. Multiple Series. Finite sums can have more than one index of summation.

For instance,

$$\sum_{i=1}^{2} \sum_{j=1}^{3} a_{ij} = a_{11} + a_{12} + a_{13} + a_{21} + a_{22} + a_{23},$$

$$\sum_{i=1}^{4} \sum_{j=1}^{i} \frac{1}{ij} = \frac{1}{1^1} + \frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{3^1} + \frac{1}{3^2} + \frac{1}{3^3} + \frac{1}{4^1} + \frac{1}{4^2} + \frac{1}{4^3} + \frac{1}{4^4}$$

and the like (compare with Sec. XVI.8).

Infinite series can also have more than one summation index; these series are called *double*, *triple* etc. and, generally, *multiple*. We shall restrict ourselves to investigating a double series of the simplest form

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} \tag{31}$$

The series of higher multiplicity and also the series having variable summation indices in the inner sum (the latter finite sum belongs to this type) are treated similarly.

Let us first suppose that all $a_{ij} \geqslant 0$. We arrange all the terms of series (31) in an arbitrary order and form an ordinary series. For instance, we can arrange them as follows:

$$a_{11} + a_{12} + a_{21} + a_{13} + a_{22} + a_{31} + a_{14} + a_{23} + $$
$$+ a_{32} + a_{41} + \ldots \tag{32}$$

It is the sum of this series (which does not depend on the order of the summands; see property 4 in Sec. 3) that is called the sum of series (31). There can be two cases here, the convergence and the divergence, i.e.

$$\text{either } \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} < \infty \quad \text{or} \quad \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \infty$$

Hence, for $a_{ij} \geqslant 0$, the sum of series (31) is independent of the order of summation but of course we must not omit a single term when forming a series of type (32). In particular, we can perform the summation in the following ways:

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_{ij} \right) = \sum_{j=1}^{\infty} \left(\sum_{i=1}^{\infty} a_{ij} \right) \tag{33}$$

If the terms a_{ij} are of any sign or are complex numbers, the series satisfying the condition

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |a_{ij}| < \infty \tag{34}$$

represent the simplest case in which we call the series *absolutely convergent*. If condition (34) holds the series (31) also converges and its sum can be computed by means of any formula of form (32) or (33) or with the help of the formula

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \lim_{M,\, N \to \infty} \sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij}$$

and the like. If condition (34) is violated the result of a summation of series (31) may depend on the order of the terms (see Sec. 3), and then we have a more complicated case.

In particular, a double series is obtained when we multiply two absolutely convergent series

$$S_1 = \sum_{i=1}^{\infty} a_i \quad \text{and} \quad S_2 = \sum_{i=1}^{\infty} b_i = \sum_{j=1}^{\infty} b_j$$

Before performing the multiplication we have changed the notation of the summation index in one of the series. The multiplication can be performed as follows:

$$S_1 S_2 = \sum_{i=1}^{\infty} a_i \sum_{j=1}^{\infty} b_j = \sum_{i=1}^{\infty} \left(a_i \sum_{j=1}^{\infty} b_j \right) = \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_i b_j \right) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_i b_j$$

When writing the last equality we have applied formula (33) because the series are absolutely convergent. Thus, series of this type are multiplied according to the rule of multiplication of finite sums

(i.e. each term of the first series is multiplied by each term of the second series and so on) which results in an absolutely convergent double series.

An analogous result is obtained when we multiply a greater number of absolutely convergent series.

§ 2. Functional Series

7. **Deviation of Functions.** If the terms of a series are not numbers (as in § 1) but functions there appears a question in what sense we must understand the fact that the partial sums which are functions approach the sum of the series (which is also a function). in the case of convergence. Hence. the question is how to estimate the deviation of one function from another.

It turns out that we can do this in different ways which are not equivalent to one another whereas the deviation of two numbers c and b is always characterized by the quantity $|a - b|$.

Let us be given two functions $f(x)$ and $\varphi(x)$ defined in the same finite interval $a \leqslant x \leqslant b$. The quantity

$$\max_{c \leqslant x \leqslant b} |f(x) - \varphi(x)| \qquad (35)$$

is called the maximal (uniform) deviation of the functions f and φ from each other. It is also sometimes referred to as Chebyshev's deviation. The geometric meaning of the quantity is illustrated in Fig. 336. This notion can be applied only to bounded functions. As a rule. we use it when continuous functions are considered. If the uniform deviation of two functions is small the difference between the values of $f(x)$ and $\varphi(x)$ is small at each point x of the interval $c \leqslant x \leqslant b$ and vice versa.

The quantity

$$\int_c^b |f(x) - \varphi(x)|\, dx \qquad (36)$$

is called the mean deviation of the functions f and φ from each other. Its geometric meaning is implied by Fig. 336: the quantity equals the area shaded in the figure (taken in its absolute value). The so-called mean square deviation is defined as

$$\sqrt{\int_c^b [f(x) - \varphi(x)]^2\, dx} \qquad (37)$$

which is analogous to (36) in many respects but is more convenient for calculations. Deviations (36) and (37) are used not only for

ntinuous functions but also for unbounded ones if the integral
vhich is improper in this case) is convergent (Sec. XIV.16). There
·e some other forms of deviation which are also of use but we do
ot discuss them here.

If we replace the difference $f(x) - \varphi(x)$ entering into formulas .
6) and (37) by maximal deviation (35) the integrals can only
ιcrease and hence we obtain

$$\int_a^b |f(x) - \varphi(x)| \, dx \leqslant (b-a) \max_{a \leqslant x \leqslant b} |f(x) - \varphi(x)|$$

$$\text{and} \quad \sqrt{\int_a^b [f(x) - \varphi(x)]^2 \, dx} \leqslant \sqrt{b-a} \max_{a \leqslant x \leqslant b} |f(x) - \varphi(x)| \tag{38}$$

onsequently, if the maximal deviation of two functions from each
ther is small their mean and mean square deviations are also
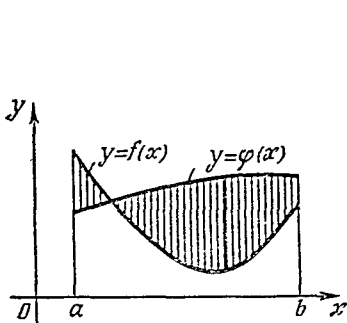
· Fig. 336
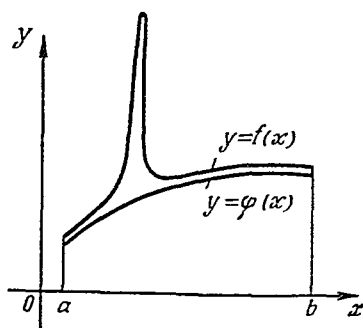
Fig. 337

mall. At the same time it can happen that the maximal deviation
f two functions is large whereas their mean deviations are small.
The possibility is illustrated in Fig. 337.

8. Convergence of a Functional Series. We now consider a series
of the form

$$f_1(x) + f_2(x) + \ldots + f_n(x) + \ldots \tag{39}$$

vhose terms are the functions $f_k(x)$ $(k = 1, 2, \ldots)$ defined over
he same finite interval $a \leqslant x \leqslant b$.

We say that the series converges to a function $S(x)$ on this inter-
val which is called the sum of the series if the deviation of the par-
ial sum $S_n(x) = \sum_{k=1}^{n} f_k(x)$ from $S(x)$ tends to zero as n increases.

Depending on the form of the deviation (see Sec. 7) we speak about

the type of the convergence of series (39). For instance, if

$$\max_{a \leqslant x \leqslant b} |S(x) - S_n(x)| \xrightarrow[n \to \infty]{} 0$$

we say that series (39) **uniformly converges** to its sum $S(x)$. Similarly, we say that the series **converges to $S(x)$ in the mean** or **in the mean square*** depending on whether we have

$$\int_a^b |S(x) - S_n(x)| \, dx \xrightarrow[n \to \infty]{} 0$$

or

$$\sqrt{\int_a^b [S(x) - S_n(x)]^2 \, dx} \xrightarrow[n \to \infty]{} 0$$

Inequalities (38) indicate that if series (39) converges uniformly it also converges in the mean (of any order) to the same sum. The converse statement may not be true in the general case.

If series (39) uniformly converges to the sum $S(x)$ on an interval $a \leqslant x \leqslant b$ we have

$$f_1(c) + f_2(c) + \ldots + f_n(c) + \ldots = S(c)$$

for each number c belonging to the interval. Actually, the difference between the value of the nth partial sum of the series at the point c and $S(c)$ does not exceed the maximal deviation of $S_n(x)$ from $S(x)$ (why?) and therefore it tends to zero as $n \to \infty$. This property makes it possible to obtain numerical series with known sums from a functional series when its sum is known.

To test a series for uniform convergence we usually apply Weierstrass' test whose condition is sufficient for the uniform convergence: if all $f_n(x)$ ($a \leqslant x \leqslant b$) satisfy the inequalities

$$|f_n(x)| \leqslant a_n \quad (n = 1, 2, 3, \ldots; a \leqslant x \leqslant b) \quad \text{and} \quad \sum_{n=1}^{\infty} a_n < \infty \quad (40)$$

---

* Generally, we say that a series with partial sums $S_n(x)$ converges in the mean of order p to its sum $S(x)$ on an interval $a \leqslant x \leqslant b$ if

$$\lim_{n \to \infty} \left(\int_a^b |S(x) - S_n(x)|^p \, dx \right)^{\frac{1}{p}} = 0$$

Hence, in this English edition the term "convergence in the mean" is understood as "convergence in the mean of order one" and the term "convergence in the mean square" as "convergence in the mean of order two". When the term "convergence in the mean" is used in mathematical books without qualification it is sometimes understood as "convergence in the mean of order two" and sometimes as "convergence in the mean of order one".—*Tr.*

series (39) converges uniformly. To prove the assertion we take advantage of the comparison test (see Sec. 1) and conclude that series (39) absolutely converges (as a numerical series) to a sum $S(x)$ for each fixed value of x. At the same time

$$\max_{a \leqslant x \leqslant b} |S(x) - S_n(x)| = \max_{a \leqslant x \leqslant b} \Big| \sum_{k=n+1}^{\infty} f_k(x) \Big| \leqslant$$
$$\leqslant \max_{a \leqslant x \leqslant b} \sum_{k=n+1}^{\infty} |f_k(x)| \leqslant \sum_{k=n+1}^{\infty} a_k$$

Since the last sum is the remainder of a convergent series (see Sec. III.6) it tends to zero as $n \to \infty$.

Condition (40) can also be written as

$$\sum_{n=1}^{\infty} \max_{a \leqslant x \leqslant b} |f_n(x)| < \infty$$

because the terms of the latter series can be taken as a_n. The tests for convergence of series (39) in the mean or in the mean square are similar to the above test. The conditions which are sufficient for these types of convergence are put down, respectively, in the forms

$$\sum_{n=1}^{\infty} \int_a^b |f_n(x)| \, dx < \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \sqrt{\overline{\int_a^b [f_n(x)]^2 \, dx}} < \infty$$

but we shall not give the proof here.

There are cases when series (39) is divergent on the whole interval $a \leqslant x \leqslant b$ but convergent on some subinterval $a_1 \leqslant x \leqslant b_1$ for which $a \leqslant a_1 < b_1 \leqslant b$. The interval $a_1 \leqslant x \leqslant b_1$ is called the **domain of convergence** of series (39).

In conclusion we note that an arbitrary variation of a finite number of terms of series (39) does not affect its convergence or divergence (although this can change its sum). This property is analogous to the corresponding property of number series (see Sec. III.6).

9. Properties of Functional Series.

1. The sum of a uniformly convergent series whose terms are continuous functions cannot have discontinuities. Indeed, if

$$f_1(x) + f_2(x) + \ldots + f_n(x) + \ldots = S(x) \tag{41}$$
$$(a \leqslant x \leqslant b)$$

we have

$$S(x) = [f_1(x) + \ldots + f_n(x)] + [f_{n+1}(x) + f_{n+2}(x) + \ldots] =$$
$$= S_n(x) + R_n(x) \tag{42}$$

If the terms of the series are continuous functions, $S_n(x)$ is also continuous as a sum of a finite number of continuous functions (see

Sec. III.14). On the other hand, if series (41) is uniformly conver-
gent its remainder $R_n(x)$ will be arbitrarily small over the whole
interval $a \leqslant x \leqslant b$ for sufficiently large values of n. Therefore
a small variation of x yields small variations both of $S_n(x)$ and
$R_n(x)$, and thus the whole sum (42) gains a small increment as well
which means that the sum cannot have discontinuities.

We sometimes consider series of form (41) on a finite or infinite
interval $a < x < b$ which uniformly converge not on the whole
interval but only on each proper subinterval $a_1 \leqslant x \leqslant b_1$ lying
entirely in the interior of the former interval. Then we can apply
the above property to the interval $a_1 \leqslant x \leqslant b_1$ and then, making
a_1 approach a and b_1 approach b, conclude that the sum of the series
cannot have discontinuities in the original interval $a < x < b$.
Analogous conclusions are also true for the properties enumera-
ted below.

If the terms of a series of form (41) are discontinuous we can apply
the above argument and conclude that if series (41) converges uni-
formly its sum can have discontinuities only at the points where the
terms are discontinuous. In contrast to it, if a series converges in
the mean its sum can have new discontinuities and, moreover, it
can be discontinuous even when all the summands are continuous
functions. This is connected with the fact that continuous functions
$S_n(x)$ can converge in the mean to a discontinuous function, as it
is illustrated in Fig. 270.

2. A uniformly convergent series can be *integrated termwise*,
i.e. under this assumption (41) implies, for any x_0 and x from the
interval $a \leqslant x \leqslant b$, that

$$\int_{x_0}^{x} f_1(t)\,dt + \int_{x_0}^{x} f_2(t)\,dt + \ldots + \int_{x_0}^{x} f_n(t)\,dt + \ldots = \int_{x_0}^{x} S(t)\,dt$$

where the series thus obtained is uniformly convergent on the inter-
val $a \leqslant x \leqslant b$. In fact, we have

$$\left| \int_{x_0}^{x} S(t)\,dt - \sum_{h=1}^{n} \int_{x_0}^{x} f_h(t)\,dt \right| = \left| \int_{x_0}^{x} \left[S(t) - \sum_{h=1}^{n} f_h(t) \right] dt \right| =$$

$$= \left| \int_{x_0}^{x} [S(t) - S_n(t)]\,dt \right| \leqslant \left| \int_{x_0}^{x} |S(t) - S_n(t)|\,dt \right| \leqslant$$

$$\leqslant \int_{a}^{b} |S(t) - S_n(t)|\,dt \leqslant (b-a) \cdot \max_{a \leqslant t \leqslant b} |S(t) - S_n(t)| \xrightarrow[n \to \infty]{} 0$$

In the case of convergence in the mean we can write the same
inequalities omitting the last one and thus prove that a series con-
vergent in the mean can be integrated term-by-term and that the

series obtained after the integration uniformly converges on the interval $a \leqslant x \leqslant b$.

3. A series with continuous terms can be *differentiated termwise* if this results in a uniformly convergent series, that is under these assumptions (41) implies

$$f_1'(x) + f_2'(x) + \ldots + f_n'(x) + \ldots = S'(x)$$

To prove the property we denote the sum of the latter series by $Q(x)$. Then integrating this series term-by-term (which is permissible on the basis of property 2) we arrive at the equality

$$S(x) - S(x_0) = \int_{x_0}^{x} Q(t)\, dt$$

Finally, differentiating the last relation, we obtain $Q(x) = S'(x)$ which is what we set out to prove.

It is possible to specify the notion of convergence of a functional series with the help of generalized functions (see Sec. XIV.27) and then all the restrictions imposed on the character of the convergence may be dropped and we can integrate and differentiate termwise any convergent (in the generalized sense) series.

§ 3. Power Series

10. Interval of Convergence. A power series is written in the form

$$a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots \tag{43}$$

We have already encountered series of this type in our course (see Sec. IV.16). Now we proceed to give the general theory of such series. For the sake of simplicity let us suppose that there exists a finite or infinite limit of the form

$$\lim_{n \to \infty} \frac{|a_n|}{|a_{n+1}|} = R \tag{44}$$

although the final results of our investigation will be valid for the general case.

We can easily find out for what numerical values of x series (43) converges. Since

$$\lim_{n \to \infty} \frac{|a_{n+1} x^{n+1}|}{|a_n x^n|} = \lim_{n \to \infty} \frac{|x|}{\left|\frac{a_n}{a_{n+1}}\right|} = \frac{|x|}{|R|} \tag{45}$$

we conclude, on the basis of D'Alembert's test (see Sec. 2), that series (43) is absolutely convergent for $|x| < R$, i.e. for

$$-R < x < R \tag{46}$$

Interval (46) is referred to as the **interval of convergence** of power series (43), and R is called the **radius of convergence**. Series (43) diverges for $|x| > R$, that is for $-\infty < x < -R$ and $R < x < \infty$. Indeed, outside the interval of convergence limit (45) exceeds unity which implies the divergence. Limit (45) is equal to 1 for $x = \pm R$, that is D'Alembert's test is inapplicable to the end-points of the interval of convergence. Examples show that depending on the particular properties of a power series it can be convergent or divergent at an end-point of its interval of convergence.

Limit (44) does not exist in some cases but even then the interval of convergence can sometimes be found by means of D'Alembert's test. As an example, take the series

$$1 - \frac{x^3}{2 \cdot 2^2} + \frac{x^6}{3 \cdot 2^4} - \frac{x^9}{4 \cdot 2^6} + \frac{x^{12}}{5 \cdot 2^8} - \cdots \qquad .$$

Limit (44) does not exist for the series (why?). The series converges for the values of x which yield

$$\lim_{n \to \infty} \left\{ \frac{|x^{3(n+1)}|}{(n+2)\, 2^{2(n+1)}} : \frac{|x^{3n}|}{(n+1)\, 2^{2n}} \right\} = \frac{|x|^3}{2^2} < 1$$

Hence, the series converges for $|x^3| < 2^2 = 4$, that is its interval of convergence is $-\sqrt[3]{4} < x < \sqrt[3]{4}$. The series diverges at the end-point $x = -\sqrt[3]{4}$ of the interval of convergence and conditionally converges at the end-point $x = \sqrt[3]{4}$ (check up these assertions!).

If D'Alembert's test is inapplicable we can nevertheless prove that the domain of convergence of a series of form (43) is an interval of type (46) but it is more difficult to find the value of R in this case.

If $R = \infty$ series (43) converges for all x, i.e. over the whole x-axis, although it can be inapplicable for practical calculations when the values of $|x|$ become large (see the end of Sec. 4). The case $R = 0$ is also theoretically possible. But then series (43) converges only at the single point $x = 0$ and therefore we shall not treat such series here.

Let the reader verify that the radii of convergence for series (IV.55)-(IV.61) are equal, respectively, to ∞, ∞, ∞, ∞, ∞, 1 and 1.
We also consider power series of the form

$$a_0 + a_1 (x - a) + a_2 (x - a)^2 + \ldots + a_n (x - a)^n + \ldots \qquad (47)$$

Denoting $x - a$ by x_1 we reduce the series to form (43) (with respect to the new variable x_1). Hence, the series converges for

$$-R < x - a < R, \quad \text{i.e. for} \quad a - R < x < a + R$$

11. Properties of Power Series.
1. A series of the form

$$a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots \qquad (48)$$

uniformly converges (see Sec. 8) on each interval $-R_1 \leqslant x \leqslant R_1$ where $0 < R_1 < R$ and R is the radius of convergence of series (48). Indeed, we can write

$$|a_0| = |a_0|, \quad |a_1 x| \leqslant |a_1 R_1|, \quad |a_2 x^2| \leqslant |a_2 R_1^2|,$$
$$|a_3 x^3| \leqslant |a_3 R_1^3|, \ldots$$

for such an interval and hence the terms of series (48) do not exceed the corresponding terms of the series

$$|a_0| + |a_1 R_1| + |a_2 R_1^2| + |a_3 R_1^3| + \ldots$$

in their absolute values. The latter series converges since R_1 lies inside the interval of convergence. Thus, by Weierstrass' test (see Sec. 8), series (48) uniformly converges on the interval.

In the general case a power series may not uniformly converge on the entire interval of convergence. But Abel proved that if series (48) converges at an end-point of the interval of convergence, the interval on which the uniform convergence is guaranteed can be extended to this end-point.

2. The sum of series (48) is continuous inside its interval of convergence. Indeed, this follows from property 1 in Sec. 9. Besides, Abel's theorem mentioned above implies that if series (48) converges at an end-point of the interval of convergence its sum is continuous at the end-point.

3. Term-by-term differentiation or integration of series (48) does not change its radius of convergence. For instance, integrating series (48) termwise we obtain the series

$$a_0 x + \frac{a_1}{2} x^2 + \frac{a_2}{3} x^3 + \ldots + \frac{a_{n-1}}{n} x^n + \frac{a_n}{n+1} x^{n+1} + \ldots$$

Let us compute its radius of convergence:

$$\lim_{n \to \infty} \frac{\frac{|a_{n-1}|}{n}}{\frac{|a_n|}{n+1}} = \lim_{n \to \infty} \frac{(n+1)}{n} \frac{|a_{n-1}|}{|a_n|} = \lim_{n \to \infty} \frac{n+1}{n} \lim_{n \to \infty} \frac{|a_{n-1}|}{|a_n|} = 1 \cdot R$$

Thus we obtain the same value of the radius, see (44).

4. The relation

$$a_0 + a_1 x + \ldots + a_n x^n + \ldots = S(x) \quad (-R < x < R)$$

can be termwise integrated and differentiated any number of times. This follows from properties 1 and 3 proved above and properties 2 and 3 in Sec. 9 because if the radius does not change when we integrate or differentiate once it cannot change when the operations are performed repeatedly.

In particular, properties 4 and 2 imply that the sum of a power series possesses continuous derivatives of all the orders within its interval of convergence.

For example, let us take the series

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + x^8 - \ldots \quad (-1 < x < 1)$$

which can be obtained from series (IV.60) by putting $a = -1$ or by applying formula (III.7) for the sum of an infinite geometric progression (with common ratio less than unity in its modulus). Integrating termwise we obtain

$$\int_0^x \frac{1}{1+x^2}\, dx = \arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots \qquad (49)$$

$$(-1 < x < 1)$$

The series on the right-hand side being convergent for $x = 1$ as well, Abel's theorem implies the validity of formula (49) for $x = 1$. Hence we have managed to find the sum of an interesting series of the form

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \ldots = \arctan 1 = \frac{\pi}{4}$$

Performing termwise integrations and differentiations of a given series we can sometimes reduce the series to a series whose sum is known and thus find the sum of the series in question. As an example, let us find the sum of the series

$$2 + \frac{3}{1!} x + \frac{4}{2!} x^2 + \frac{5}{3!} x^3 + \ldots = S(x)$$

By D'Alembert's test, we readily conclude that the series converges throughout the whole x-axis, that is $R = \infty$. Let us multiply both sides by x and integrate the result from zero to some x:

$$\int_0^x xS(x)\, dx = x^2 + \frac{x^2}{1!} + \frac{x^4}{2!} + \frac{x^5}{3!} + \ldots =$$

$$= x^2 \left(1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \right) = x^2 e^x$$

We obtain, by differentiating,

$$xS(x) = (x^2 e^x)' = 2x e^x + x^2 e^x$$

i.e. finally we have

$$S(x) = (2 + x) e^x$$

Let us consider an example of different kind. Let it be necessary to find the sum of the series

$$\frac{x^2}{1\cdot2}+\frac{x^3}{2\cdot3}+\frac{x^4}{3\cdot4}+\frac{x^5}{4\cdot5}+\ldots=\sigma(x) \tag{50}$$

For this purpose we differentiate it termwise:

$$\sigma'(x)=\frac{x}{1}+\frac{x^2}{2}+\frac{x^3}{3}+\frac{x^4}{4}+\ldots=-\ln(1-x)$$

[see formula (IV.61)]. It follows that

$$\sigma(x)=-\int\ln(1-x)\,dx=-x\ln(1-x)-\int\frac{x}{1-x}\,dx=$$
$$=x+(1-x)\ln(1-x)+C \tag{51}$$

To determine the value of C we put $x=0$ in formulas (50) and (51). This yields $0=\sigma(0)=C$. Finally,

$$\sigma(x)=x+(1-x)\ln(1-x)$$

In other examples of this type we can encounter integrals which are inexpressible in terms of elementary functions.

The sum of a functional series can sometimes be found by forming a differential equation which is satisfied by the sum and solving it. For instance, let us find the sum of the series

$$\frac{x}{1!}+\frac{x^4}{4!}+\frac{x^7}{7!}+\frac{x^{10}}{10!}+\frac{x^{13}}{13!}+\ldots=p(x) \tag{52}$$

To do this we differentiate formula (52) three times:

$$1+\frac{x^3}{3!}+\frac{x^6}{6!}+\frac{x^9}{9!}+\frac{x^{12}}{12!}+\ldots=p'(x), \tag{53}$$

$$\frac{x^2}{2!}+\frac{x^5}{5!}+\frac{x^8}{8!}+\frac{x^{11}}{11!}+\ldots=p''(x), \tag{54}$$

$$\frac{x}{1!}+\frac{x^4}{4!}+\frac{x^7}{7!}+\frac{x^{10}}{10!}+\ldots=p'''(x)$$

We see that we have arrived at the original series, i.e.

$$p'''(x)-p(x)=0$$

Applying the methods of Sec. XV.17 and solving the equation we find:

$$p(x)=C_1e^x+e^{-\frac{x}{2}}\left(C_2\cos\frac{\sqrt{3}}{2}x+C_3\sin\frac{\sqrt{3}}{2}x\right) \tag{55}$$

To compute C_1, C_2 and C_3 we substitute $x=0$ into formulas (52), (53) and (54) which results in $p(0)=0$, $p'(0)=1$ and $p''(0)=0$. These values determine the initial conditions for $p(x)$. By formula

(55), we obtain

$$C_1 + C_2 = 0, \quad C_1 - \frac{1}{2}C_2 + \frac{\sqrt{3}}{2}C_3 = 1, \quad C_1 - \frac{1}{2}C_2 - \frac{\sqrt{3}}{2}C_3 = 0$$

which implies

$$C_1 = \frac{1}{3}, \quad C_2 = -\frac{1}{3}, \quad C_3 = \frac{1}{\sqrt{3}}$$

Finally, the sum of series (52) is expressed as

$$\frac{x}{1!} + \frac{x^4}{4!} + \frac{x^7}{7!} + \frac{x^{10}}{10!} + \ldots = \frac{1}{3}e^x + e^{-\frac{x}{2}}\left(-\frac{1}{3}\cos\frac{\sqrt{3}}{2}x + \frac{1}{\sqrt{3}}\sin\frac{\sqrt{3}}{2}x\right)$$

In other cases we can sometimes similarly reduce the sum of a given number series to an integral or even to a simple combination of mathematical constants (i.e. to integers, numbers π and e etc.) and functions of the constants. As an example, let us take the sum

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots = S \tag{56}$$

To compute it, we introduce an auxiliary series of the form

$$\frac{x}{1^2} + \frac{x^2}{2^2} + \frac{x^3}{3^2} + \ldots = q(x) \quad (-1 \leqslant x \leqslant 1)$$

Differentiating the series we derive

$$q(x) = -\int_0^x \frac{\ln(1-x)}{x}\,dx$$

(check it up!). Substituting $x = 1$ we obtain the sum of series (56):

$$S = -\int_0^1 \frac{\ln(1-x)}{x}\,dx \tag{57}$$

The corresponding indefinite integral is not an elementary function but nevertheless it is sometimes preferable to express the sum of a series in the form of a definite integral. By the way, in Sec. 25 we shall give another method of computing the sum of series (56) which will show that the sum equals $\frac{\pi^2}{6}$. From this, in particular, we obtain the numerical value of integral (57).

12. Algebraic Operations on Power Series. The power series being absolutely convergent in the interior of their intervals of convergence, we can termwise add them together, multiply by a common factor (see Sec. 3) and multiply them by one another following the rules of multiplication of polynomials (see Sec. 6).

For example, let us consider the multiplication of the power series expressing the functions e^x and $\ln{(1+x)}$:

$$e^x \ln{(1+x)} = \left(1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots\right)\left(x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots\right) =$$

$$= x + \left(\frac{1}{1!} - \frac{1}{2}\right) x^2 + \left(\frac{1}{2!} - \frac{1}{2 \cdot 1!} + \frac{1}{3}\right) x^3 +$$

$$+ \left(\frac{1}{3!} - \frac{1}{2 \cdot 2!} + \frac{1}{3 \cdot 1!} - \frac{1}{4}\right) x^4 + \left(\frac{1}{4!} - \frac{1}{2 \cdot 3!} + \frac{1}{3 \cdot 2!} - \frac{1}{4 \cdot 1!} + \frac{1}{5}\right) x^5 +$$

$$+ \ldots = x + \frac{1}{2} x^2 + \frac{1}{3} x^3 + \frac{1}{6} x^4 + \frac{3}{40} x^5 + \ldots$$

Obviously, we can compute as many coefficients entering into the product as needed. The radius of convergence of the first series is equal to ∞ whereas that of the second series is equal to 1. Hence, the above result is valid for the interval $-1 < x < 1$ in which both series are absolutely convergent.

The division of series is performed in a similar way. For instance, consider the ratio of the power series for $\sin x$ and $\cos x$. To perform the division we take these series arranged in ascending power of the variable and divide them as if they were polynomials:

$$\frac{\sin x}{\cos x} = \frac{\dfrac{x}{1!} - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \dfrac{x^7}{7!} + \ldots}{1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} - \dfrac{x^6}{6!} + \ldots} = \frac{\dfrac{x}{1} - \dfrac{x^3}{6} + \dfrac{x^5}{120} - \dfrac{x^7}{5040} + \ldots}{1 - \dfrac{x^2}{2} + \dfrac{x^4}{24} - \dfrac{x^6}{720} + \ldots} ;$$

$$
\begin{array}{r}
x + \dfrac{x^3}{3} + \dfrac{2}{15} x^5 + \dfrac{17}{315} x^7 + \ldots \\
1 - \dfrac{x^2}{2} + \dfrac{x^4}{24} - \dfrac{x^6}{720} \Big|\ \overline{\ x - \dfrac{x^3}{6} + \dfrac{x^5}{120} - \dfrac{x^7}{5040} + \ldots\ } \\
\underline{x - \dfrac{x^3}{2} + \dfrac{x^5}{24} - \dfrac{x^7}{720} + \ldots} \\
\dfrac{x^3}{3} - \dfrac{x^5}{30} + \dfrac{x^7}{840} + \ldots \\
\underline{\dfrac{x^3}{3} - \dfrac{x^5}{6} + \dfrac{x^7}{720} + \ldots} \\
\dfrac{2x^5}{15} - \dfrac{4x^7}{315} + \ldots \\
\underline{\dfrac{2x^5}{15} - \dfrac{x^7}{15} + \ldots} \\
\dfrac{17}{315} x^7 + \ldots \\
\underline{\dfrac{17}{315} x^7 + \ldots} \\
\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot
\end{array}
$$

Thus, the expansion of the tangent into a power series is of the following form:

$$\tan x = \frac{\sin x}{\cos x} = x + \frac{x^3}{3} + \frac{2}{15} x^5 + \frac{17}{315} x^7 + \dots \qquad (58)$$

To compute the subsequent terms in (58) we must put down the subsequent terms in the expansions of $\sin x$ and $\cos x$ and continue the division. It can be proved that formula (58) is valid for $|x| < < \frac{\pi}{2}$.

Expansion (58) can also be obtained by means of the method of undetermined coefficients. To apply the method we note that $\tan x$ is an odd function and therefore its expansion must contain only the odd powers of x:

$$\tan x = a_1 x + a_3 x^3 + a_5 x^5 + a_7 x^7 + \dots$$

But $\cos x \cdot \tan x = \sin x$, and therefore

$$\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots\right)(a_1 x + a_3 x^3 + a_5 x^5 + a_7 x^7 + \dots) =$$

$$= \frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

Removing the brackets and equalling the coefficients in like powers of x we obtain:

$$a_1 = \frac{1}{1!}, \quad a_3 - \frac{a_1}{2!} = -\frac{1}{3!}, \quad a_5 - \frac{a_3}{2!} + \frac{a_1}{4!} = \frac{1}{5!},$$

$$a_7 - \frac{a_5}{2!} + \frac{a_3}{4!} - \frac{a_1}{6!} = -\frac{1}{7!}, \quad \dots$$

Hence, the coefficients $a_1, a_3, a_5, \dots$ can be found in succession from the above relations.

Another method of manipulating series is the substitution of a series into a series. For instance, we can substitute a series of the form

$$y = f(x) = a_0 + a_1 x + a_2 x^2 + \dots$$

into the series

$$\varphi(y) = b_0 + b_1 y + b_2 y^2 + \dots \qquad (59)$$

or, in a more general case, into the series

$$\psi(y) = c_0 + c_1(y - a) + c_2(y - a)^2 + \dots$$

For instance, taking series (59) we obtain the expression

$$\varphi\,(f\,(x)) = b_0 + b_1\,(a_0 + a_1 x + a_2 x^2 + \ldots) + \\ + b_2\,(a_0 + a_1 x + a_2 x^2 + \ldots)^2 + \ldots$$

in which we must remove the brackets and combine similar terms. For the result to be valid for $x = 0$ it is necessary that series (59) converge for $y = a_0$ (why is it so?). This means that a_0 must lie within the interval of convergence of series (59). By the way, if the condition is not fulfilled the calculations themselves will indicate the mistake.

Take an example:

$$\ln\,(1 + \sin x) = \frac{\sin x}{1} - \frac{(\sin x)^2}{2} + \frac{(\sin x)^3}{3} - \ldots =$$

$$= \frac{x - \dfrac{x^3}{6} + \dfrac{x^5}{120} - \dfrac{x^7}{5040} + \ldots}{1} - \frac{\left(x - \dfrac{x^3}{6} + \dfrac{x^5}{120} - \ldots\right)^2}{2} +$$

$$+ \frac{\left(x - \dfrac{x^3}{6} + \dfrac{x^5}{120} - \ldots\right)^3}{3} - \ldots = \frac{x - \dfrac{x^3}{6} + \dfrac{x^5}{120} - \dfrac{x^7}{5040} + \ldots}{1} -$$

$$- \frac{x^2 - \dfrac{x^4}{3} + \dfrac{2}{45}x^6 + \ldots}{2} + \frac{x^3 - \dfrac{x^5}{2} + \dfrac{13}{120}x^7 + \ldots}{3} -$$

$$- \frac{x^4 - \dfrac{2}{3}x^6 + \ldots}{4} + \frac{x^5 - \dfrac{5}{6}x^7 + \ldots}{5} - \frac{x^6 + \ldots}{6} + \frac{x^7 + \ldots}{7} - \ldots$$

When calculating in succession the powers entering into the resulting series we have performed the multiplications for the Taylor series of $\sin x$ according to the rule of multiplication of polynomials and dropped the powers of x higher than the power corresponding to the desired accuracy of calculations (higher than x^8). Now, combining similar terms we finally derive

$$\ln\,(1 + \sin x) = x - \frac{x^2}{2} + \frac{x^3}{6} - \frac{x^4}{12} + \frac{x^5}{24} - \frac{x^6}{45} + \frac{61x^7}{5040} - \ldots$$

The methods discussed in Secs. 11, 12 make it possible to obtain the expansions of many other functions by taking advantage of the simplest series given in Sec. IV.16. It is sometimes difficult to write down the explicit expression of the general term of such an expansion but at the same time we can always find any number of terms which is usually sufficient for practical purposes.

13. **Power Series as a Taylor Series.** We now consider the sum of a series of the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots \qquad (60)$$

$$(-R < x < R)$$

The coefficients of the series can be easily expressed in terms of its sum. For this purpose we perform successive differentiations of formula (60) and substitute $x = 0$ into the results, as it was done in Sec. IV.15. This yields

$$f(0) = a_0;$$

$$f'(x) = 1a_1 + 2a_2 x + 3a_3 x^2 + \ldots, \qquad f'(0) = 1a_1;$$

$$f''(x) = 1 \cdot 2a_2 + 2 \cdot 3a_3 x + 3 \cdot 4a_4 x^2 + \ldots, \qquad f''(0) = 1 \cdot 2a_2;$$

$$f'''(x) = 1 \cdot 2 \cdot 3a_3 + 2 \cdot 3 \cdot 4a_4 x + 3 \cdot 4 \cdot 5a_5 x^2 + \ldots, \qquad f'''(0) = 1 \cdot 2 \cdot 3a_3 \text{ etc.}$$

Finding a_0, a_1, a_2, ... from these relations and substituting them into (60) we obtain the expression

$$f(x) = f(0) + \frac{f'(0)}{1!} x + \frac{f''(0)}{2!} x^2 + \frac{f'''(0)}{3!} x^3 + \ldots$$

$$\ldots + \frac{f^{(n)}(0)}{n!} x^n + \ldots \qquad (-R < x < R) \qquad (61)$$

which is nothing but Taylor's series (IV.54) we have already dealt with. Thus, *a power series is Taylor's series of its sum.*

The coefficients of series (61) being uniquely expressed in terms of its sum, we can assert, in particular, that if the sums of two power series are identically equal their coefficients in like powers of x also coincide. Accordingly, if the sum of a power series is identically equal to zero all its coefficients are also equal to zero.

In the above argument we have regarded a power series as being given. But in practice we usually deal with the reverse problem of expanding a given function $f(x)$. Then, naturally, we encounter the problem of determining the range of the values of x for which formula (61) is valid. On the basis of Secs. IV.15, 16, we conclude that this is equivalent to the question what are the values of x for which the remainder of the corresponding finite Taylor formula tends to zero when the number n increases.

An exhaustive investigation of the remainder can be carried out only in some simple cases. Fortunately, we can easily do without such an investigation when we deal with elementary functions because it is possible to prove that formula (61) is valid for an elementary function $f(x)$ on every interval in which the series converges provided that the direct substitution of $x = 0$ into the expressions $f(x)$, $f'(x)$, $f''(x)$, ... yields the finite results $f(0)$, $f'(0)$, $f''(0)$, ... This implies, in particular, that the expansions given

in Sec. IV.16 are valid for the intervals of convergence of the corresponding series.

It should be remarked that there are many functions that cannot be expanded into power series (i.e. into Taylor's series). For instance, a function which is discontinuous in an interval or has a derivative of the first or of a higher order with a discontinuity on the interval cannot be represented by a power series. The same is true for a function which is represented by means of different formulas on different parts of the interval under consideration.

All that has been said here is directly extended to series in powers of $x - a$ of form (47) and to corresponding Taylor's series (IV.53).

14. Power Series with Complex Terms. These series are of the form

$$a_0 + a_1 z + a_2 z^2 + \ldots + a_n z^n + \ldots, \quad z = x + iy \qquad (62)$$

where the coefficients a_n and the independent variable z can assume any complex values. The theory of these series is analogous to that
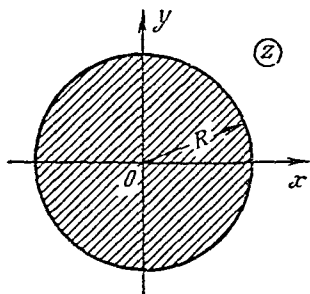
Fig. 338

of real power series but the inequality $|z| < R$ defining the values of z for which series (62) is convergent represents not an interval but a circle in the complex z-plane. The circle is referred to as the **circle of convergence** of series (62) (see Fig. 338). Similarly, for a series in powers of $z - a$ where a is a complex number the domain of convergence is specified by an equality of the form $|z - a| < R$ which defines a circle of radius R with centre at the point a. If $R = \infty$ the series converges throughout the whole

complex plane. The properties enumerated in Secs. 11 and 12 can be transferred to the series of form (62) without essential changes. The sum $S(z)$ of such a series is a complex function of a complex variable (see Sec. VIII.11). For these functions the notion of an integral is introduced by means of the notion of an antiderivative. In Sec. VIII.4 we considered several examples of defining a function for complex values of its argument by means of series of form (62).

As it was mentioned in Sec. VIII.4, the identities which are satisfied by functions for real values of the argument remain valid for its complex values when we continue the functions in the complex plane by means of the corresponding power series. To illustrate what has been said we take the equality

$$e^{\ln(1+x)} = 1 + x \qquad (63)$$

It is valid for real x, by the definition of the logarithm. Hence, if we substitute the power series of $y = \ln(1+x)$ (see Sec. 12) into the series of e^y and perform the corresponding identical transformations (removing the brackets etc.) we must obtain $1 + x$. Performing the same operations for the complex values of the argument, i.e. writing z in place of x, we obtain

$$e^{\ln(1+z)} = 1 + z \qquad (64)$$

which is what we set out to prove.

This formula implies, in particular, that the definition of the logarithm by means of power series (IV.61) into which z is substituted for x is coherent with the definition of the logarithm of a complex number given in Sec. VIII.5. The sum of the series represents only one branch of the infinite-valued function $\ln(1+z)$, namely the branch which yields zero when $z = 0$ is substituted into the function.

Formula (VIII.7) and many other formulas can be proved in a similar way.

15. Bernoullian Numbers. The so-called **Bernoullian numbers** discovered by Jacob Bernoulli are widely applied in the theory of series and particularly in the theory of power series. These numbers, which we denote by β_1, β_2, β_3, β_4, ..., are defined by a symbolic recurrence relation of the form

$$(\beta+1)^{n+1} - \beta^{n+1} = 0 \qquad (n = 1, 2, 3, \ldots)$$

in which β^k should be replaced by β_k after the brackets have been removed on the left-hand side.

For instance, putting $n = 1$, $n = 2$, $n = 3$, in succession, we obtain, respectively,

$$\beta_2 + 2\beta_1 + 1 - \beta_2 = 0$$

which yields $\beta_1 = -\dfrac{1}{2}$,

$$\beta_3 + 3\beta_2 + 3\beta_1 + 1 - \beta_3 = 0$$

which yields $\beta_2 = -\dfrac{3\beta_1 + 1}{3} = \dfrac{1}{6}$, and

$$\beta_4 + 4\beta_3 + 6\beta_2 + 4\beta_1 + 1 - \beta_4 = 0$$

which yields $\beta_3 = -\dfrac{6\beta_2 + 4\beta_1 + 1}{4} = 0$.

The subsequent calculations result in

$$\beta_4 = -\frac{1}{30}, \quad \beta_5 = 0, \quad \beta_6 = \frac{1}{42}, \quad \beta_7 = 0, \quad \beta_8 = -\frac{1}{30},$$

$$\beta_9 = 0, \quad \beta_{10} = \frac{5}{66}, \quad \beta_{11} = 0, \quad \beta_{12} = -\frac{691}{2730}, \quad \beta_{13} = 0, \quad \beta_{14} = \frac{7}{6}, \ldots$$

It is possible to prove that all β_n with odd numbers $n \geqslant 3$ are equal to zero. Introducing the notation

$$B_n = (-1)^{n-1} \beta_{2n} \quad (n = 1, 2, 3, \ldots)$$

we write

$$B_1 = \frac{1}{6}, \quad B_2 = \frac{1}{30}, \quad B_3 = \frac{1}{42}, \quad B_4 = \frac{1}{30}, \quad B_5 = \frac{5}{66}, \quad B_6 = \frac{691}{2730}, \ldots$$

These numbers are also referred to as **Bernoullian numbers**.

We now put down some formulas involving Bernoullian numbers. For the function $\zeta(x)$ (see Sec. 4) we have

$$\sum_{n=1}^{\infty} \frac{1}{n^{2k}} = \zeta(2k) = \frac{B_h (2\pi)^{2k}}{2 (2k)!} \quad (k = 1, 2, \ldots) \tag{65}$$

In particular, this yields

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{B_1 (2\pi)^2}{2 \cdot 2!} = \frac{\pi^2}{6}, \quad \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{B_2 (2\pi)^4}{2 \cdot 4!} = \frac{\pi^4}{90} \text{ etc.}$$

Formula (65) shows that all the numbers B_n are positive. Further, we have

$$\tan x = \sum_{n=1}^{\infty} \frac{2^{2n} (2^{2n} - 1)}{(2n)!} B_n x^{2n-1}$$

[compare with formula (58)] etc.

16. Applying Series to Solving Difference Equations. A **difference equation** connects an unknown quantity and its finite differences (see Sec. V.7). We first turn to the case when the unknown quantity is represented by a sequence $a_0, a_1, a_2, \ldots, a_n$. A difference equation defining the sequence can be put down in the general form

$$f(n, a_n, \Delta a_n, \Delta^2 a_n) = 0 \quad (n = 0, 1, 2, \ldots) \tag{66}$$

if we restrict ourselves to equations of the second order. Here $\Delta a_n = a_{n+1} - a_n$ and $\Delta^2 a_n = \Delta a_{n+1} - \Delta a_n$. Substituting

$$\Delta a_n = a_{n+1} - a_n \quad \text{and} \quad \Delta^2 a_n = a_{n+2} - 2a_{n+1} + a_n$$

we reduce (66) to the form

$$\varphi(n, a_n, a_{n+1}, a_{n+2}) = 0 \quad (n = 0, 1, 2, \ldots) \tag{67}$$

In a particular case the left-hand sides of equations (66) and (67) may not contain all the arguments put down there.

To solve equation (67) we can arbitrarily set two values of the unknown quantity, for instance, a_0 and a_1. Then we find a_2 by putting $n = 0$ in (67). Further, putting $n = 1$ in (67) and substituting the value a_2 found above we determine a_3 and so on. This step-by-step procedure enables us to find any desired number of terms of the sequence a_n.

A particular case of equation (67) is a *homogeneous linear equation* with constant coefficients which is written as

$$\alpha a_n + \beta a_{n+1} + \gamma a_{n+2} = 0 \quad (n = 0, 1, 2, \ldots; \quad \alpha, \beta, \gamma = const) \quad (68)$$

The solution of equation (68) can be found in the general form by means of a method which is illustrated below. Equations of any order can be solved in a similar way.

Let us form the so-called *generating function* of the sought-for sequence whose expansion into a power series of the form

$$Q = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots$$

gives rise to the terms of the sequence as the coefficients in the series. It is only the coefficients of the series that we are interested in here and therefore we are not going to consider any numerical values of x and the question on the convergence of the series. In such a case a power series is referred to as a *formal power series*. We can easily find the product

$$(\gamma + \beta x + \alpha x^2) Q = \gamma a_0 + (\beta a_0 + \gamma a_1) x + (\alpha a_0 + \beta a_1 + \gamma a_2) x^2 + \\ + (\alpha a_1 + \beta a_2 + \gamma a_3) x^3 + \ldots$$

Equation (68) implies that all the coefficients on the right-hand side, from that in x^2 onwards, are equal to zero. Performing the division we derive

$$Q = \frac{\gamma a_0 + (\beta a_0 + \gamma a_1) x}{\gamma + \beta x + \alpha x^2} \qquad (69)$$

The values of a_0 and a_1 given, we have the ratio of two polynomials with the given coefficients on the right-hand side. It can be decomposed into partial fractions of the form $\dfrac{A}{(x-a)^\alpha}$ by applying the methods discussed in Sec. VIII.10. Each fraction can be rewritten as

$$\frac{A}{(x-a)^\alpha} = \frac{A}{(-a)^\alpha \left(1 - \dfrac{x}{a}\right)^\alpha} = \frac{B}{(1 - \gamma x)^\alpha}$$

where $B = \dfrac{A}{(-a)^\alpha}$ and $\gamma = \dfrac{1}{a}$. We have $\alpha = 1$ or $\alpha = 2$ for fraction (69) but for difference equations of higher order the values of α may be greater. These fractions are expanded into power series according to the formulas which are obtained from the formula of the sum of a geometric series by means of differentiation:

$$\frac{B}{1 - \gamma x} = B + B\gamma x + B\gamma^2 x^2 + \ldots + B\gamma^n x^n + \ldots,$$

$$\frac{B}{(1 - \gamma x)^2} = \frac{1}{\gamma} \left(\frac{B}{1 - \gamma x}\right)' =$$

$$= B + 2B\gamma x + 3B\gamma^2 x^2 + \ldots + (n+1) B\gamma^n x^n + \ldots$$

and so on. Adding together the coefficients in x^n entering into all the above series we thus determine the coefficient a_n in the series $Q = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots$ and hence equation (68) has been solved in the general form.

We suggest that the reader apply the above method to deriving the general formula of the so-called *Fibonacci numbers* $a_0 = 0$, $a_1 = 1$, $a_2 = 1$, $a_3 = 2$, $a_4 = 3$, $a_5 = 5$, $\ldots$ each of which, from the third onwards, is equal to the sum of the two preceding numbers. The sought-for expression is of the form

$$a_n = \frac{(1 + \sqrt{5})^n - (1 - \sqrt{5})^n}{\sqrt{5} \cdot 2^n}$$

We also consider difference equations in which the unknown quantity is a sought-for function $y(x)$. In this case instead of (66) and (67) we have equations of the forms

$$f(x, y, \Delta_h y, \Delta_h^2 y) = 0 \quad \text{and} \quad \varphi(x, y(x), y(x + h), y(x + 2h)) = 0 \tag{70}$$

respectively.

The latter case can be reduced to the former in which the unknown quantity is represented by a sequence. For example, let $0 \leqslant x < \infty$. We introduce the notation $a_n = y(\xi + nh)$ $(n = 0, 1, 2, \ldots)$ where ξ is a constant number lying in the interval $0 \leqslant \xi < h$. Putting $x = \xi + nh$ in equation (70) we can rewrite it as

$$\varphi(\xi + nh, a_n, a_{n+1}, a_{n+2}) = 0$$

and hence, for a constant ξ, we obtain an equation of form (67). After a_n has been found we can use the arbitrariness of the choice of ξ and thus obtain the sought-for solution $y(x)$. In particular, it follows that we can arbitrarily set the values of $y(x)$ in the interval $0 \leqslant x < 2h$ for equation (70) (why is it so?).

17. Multiple Power Series. The role of multiple power series in the theory of functions of several arguments is similar to that of ordinary power series in the theory of functions of one independent variable. For the sake of simplicity we shall restrict ourselves to double power series. The power series of higher multiplicity are treated similarly.

To put down the expression of a double series it is convenient to use the double index notation which was applied to writing a double number series in Sec. 6:

$$S(x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{mn} x^m y^n =$$
$$= a_{00} + a_{10} x + a_{01} y + a_{20} x^2 + a_{11} xy + a_{02} y^2 + \ldots \tag{71}$$

If the series is absolutely convergent the order of performing its summation does not matter.

The domain of convergence of series (71) is a region of the x, y-plane (which may coincide with the whole plane in a particular case). Such a domain can be of the form represented in Fig. 339. For any fixed y we obtain a power series in powers of x whose radius of convergence R may depend on y. i.e. $R = R(y)$. Therefore the domain of convergence is symmetric with respect to the y-axis. The symmetry with respect to the x-axis is implied by the same argument. In the case of absolute convergence of series (71), $R(y)$ is a non-increasing function of y for $y \gg 0$ (why?).

Series of the form

$$\sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{mn} (x-a)^m (y-b)^n \qquad (72)$$

are treated similarly. The domain of convergence of such a series is a plane figure with centre of symmetry at the point (a, b).

The properties of multiple power series are analogous to those

Fig. 339

of ordinary power series (see Secs. 11 and 12). In particular, multiple power series are obtained in expanding a function of several variables into Taylor's series (see Sec. XII.6):

$$f(x, y) = f(0, 0) + \frac{f'_x(0, 0)}{1!} x + \frac{f'_y(0, 0)}{1!} y + \frac{f''_{xx}(0, 0)}{2!} x^2 +$$

$$+ \frac{f''_{xy}(0, 0)}{1! 1!} xy + \frac{f''_{yy}(0, 0)}{2!} y^2 + \cdots$$

Series (72) are obtained in the same way. Multiple power series are applicable when we use the small parameter method (see Secs. V.5 and XV.27) for an equation containing several parameters. They can also be utilized in many other problems.

18. **Functions of Matrices.** Let A be a square matrix (see Sec. XI). For definiteness, let it be of the third order (the results that we shall obtain here are true for matrices of any order). In Secs. XI.2 and XI.3 we introduced such simplest functions of matrices as A^2 and A^{-1}. But in what sense should we understand an expression of the form e^A and the like? The importance of the exponential function in mathematics indicates the advisability of putting this question.
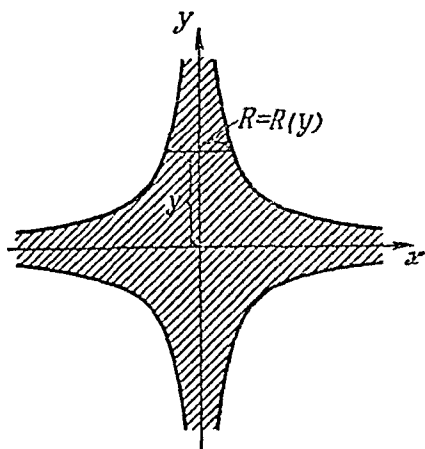
It turns out that a reasonable answer to the question can be obtained by means of power series if we apply them by analogy with Sec. VIII.4 where the notion of a function of a complex variable was defined. Suppose that we are given a function $f(x)$ which can be expanded into a power series

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_n x^n + \ldots \qquad (73)$$

By definition, we write

$$f(A) = a_0 I + a_1 A + a_2 A^2 + \ldots + a_n A^n + \ldots \qquad (74)$$

where I is the unit matrix of the same order as A. For example, we have

$$e^A = I + \frac{A}{1!} + \frac{A^2}{2!} + \frac{A^3}{3!} + \ldots + \frac{A^n}{n!} + \ldots \qquad (75)$$

The definition makes sense if series (74) converges. We can point out a simple condition for the convergence. Let us assume that series (73) has the radius of convergence R and suppose, for simplicity, that all the eigenvalues λ_1, λ_2, λ_3 of the matrix A (see Sec. XI.4) are distinct. Then, as it was shown in Sec. XI.8, the matrix A can be transformed to the diagonal form, that is there exists a non-degenerate matrix H for which

$$H^{-1}AH = \text{diag}(\lambda_1, \lambda_2, \lambda_3) = \Lambda$$

But this implies

$$A = H\Lambda H^{-1}, \quad A^2 = (H\Lambda H^{-1}) \cdot (H\Lambda H^{-1}) = H\Lambda^2 H^{-1},$$
$$A^3 = A^2 A = (H\Lambda^2 H^{-1})(H\Lambda H^{-1}) = H\Lambda^3 H^{-1}$$

etc. Consequently, series (74) can be rewritten as

$$Ha_0 I H^{-1} + Ha_1 \Lambda H^{-1} + Ha_2 \Lambda^2 H^{-1} + \ldots =$$
$$= H(a_0 I + a_1 \Lambda + a_2 \Lambda^2 + \ldots) H^{-1} \qquad (76)$$

A diagonal matrix can be easily raised to a power:

$$\Lambda^2 = \text{diag}(\lambda_1^2, \lambda_2^2, \lambda_3^2), \quad \Lambda^3 = \text{diag}(\lambda_1^3, \lambda_2^3, \lambda_3^3) \qquad (77)$$

etc.

(let the reader verify that in the general case of multiplication of diagonal matrices the product is a diagonal matrix whose elements are the products of the corresponding diagonal elements of the matrix factors). Therefore, we have

$$a_0 I + a_1 \Lambda + a_2 \Lambda^2 + \ldots =$$
$$= \text{diag}(a_0 + a_1 \lambda_1 + a_2 \lambda_1^2 + \ldots, a_0 + a_1 \lambda_2 + a_2 \lambda_2^2 + \ldots$$
$$\ldots, a_0 + a_1 \lambda_3 + a_2 \lambda_3^2 + \ldots) \qquad (78)$$

If the series forming the diagonal converge series (76) converges as well which leads to the convergence of series (74).

Thus, we can assert that if all the eigenvalues of the matrix A do not exceed R in their moduli [where R is the radius of convergence of series (73)] series (74) is convergent and even absolutely convergent. If at least one of the eigenvalues exceeds R in its absolute value series (74) is divergent. It can be proved that the above result is also valid for the case when the matrix A has multiple eigenvalues.

Formulas (76) and (78) also imply the following formula applicable to a matrix A which can be transformed to the diagonal form by means of a matrix H:

$$f(A) = H \operatorname{diag}(f(\lambda_1), f(\lambda_2), f(\lambda_3)) H^{-1}$$

What has been proved implies that series (75) is convergent for any matrix A since the corresponding series (IV.55) has the radius of convergence $R = \infty$. Another important example is the series

$$I + A + A^2 + \ldots + A^n + \ldots \qquad (79)$$

which converges if all the eigenvalues of the matrix A do not exceed unity in their moduli (why?).

Many other properties of ordinary functions can be extended to functions of matrices. These properties can be proved by manipulating series after a manner of deducing formula (64) from formula (63). For example, the identity

$$(1 + x + x^2 + \ldots)(1 - x) = \frac{1}{1-x}(1-x) = 1$$

implies the relation

$$(I + A + A^2 + \ldots)(I - A) = I$$

i.e. the sum of series (79) equals $(I - A)^{-1}$ if it is convergent. At the same time we should take into account that when proving some properties by means of series we use a permutation of factors (for instance, the relation $ab + ba = 2ab$) which may not be applicable to matrices. For instance, we apply the above relation to deducing the formula

$$e^A e^B = e^{A+B}$$

and thus it is valid for commuting matrices A and B and is inapplicable when they are non-commuting.

As an example of applying the notions introduced in this section, let us establish certain conditions guaranteeing the convergence of the iterative method of solving a system of linear algebraic equations. System (VI.19) can be rewritten in the vector form

$$x = Ax + \delta \qquad (80)$$

where δ is a given vector, A is a given coefficient matrix and x is the sought-for vector. Taking an initial approximation $x = x_0$ we obtain the successive approximations according to the iterative method:

$$x_1 = \delta + Ax_0.$$
$$x_2 = \delta + Ax_1 = \delta + A(\delta + Ax_0) = \delta + A\delta + A^2x_0,$$
$$x_3 = \delta + Ax_2 = \delta + A\delta + A^2\delta + A^3x_0$$

and so on. Generally, we have

$$x_n = (I + A + A^2 + \ldots + A^{n-1})\delta + A^n x_0 \tag{81}$$

For the process to be convergent it is necessary that the initial approximations should not affect the result obtained in the limit, that is we must have $A^n \xrightarrow[n \to \infty]{} 0$. For this to be so it is sufficient, on the basis of formulas (77), that all the eigenvalues of the matrix A be less than unity in their absolute values. It is the last condition that guarantees the convergence of the iterative method. If it is fulfilled we obtain, passing to the limit in formula (81) as $n \to \infty$, the formula

$$\bar{x} = \lim_{n \to \infty} x_n = (I + A + A^2 + \ldots + A^n + \ldots)\delta = (I - A)^{-1}\delta$$

The direct substitution of the vector $\bar{x}$ thus obtained into equation (80) shows that the vector satisfies the equation. (Perform it!)

By analogy with vector functions of a scalar argument (see Sec. VII.23), we can consider matrix functions of a scalar argument, having the form $B = B(x)$. Many of the properties of ordinary functions can be extended to this case. For instance, we often use the function

$$B = e^{Ax} \quad (-\infty < x < \infty, \ Ax = xA)$$

where A is a constant matrix. Applying the series techniques we can readily prove that $(e^{Ax})' = Ae^{Ax}$. In particular, it follows that we have $(e^{Ax}c)' = Ae^{Ax}c$ for any constant vector c. But this means that the vector function of x having the form

$$y = e^{Ax}c \tag{82}$$

is a solution of matrix equation (XV.149) with constant coefficients:

$$y' = Ay \tag{83}$$

If an initial condition of the form $y|_{x=x_0} = y_0$ is given formula (82) implies

$$y_0 = e^{Ax_0}c, \quad \text{i.e.} \quad c = e^{-Ax_0}y_0$$

Hence we obtain the following explicit formula for the solution:

$$\mathbf{y} = e^{\mathbf{A}x}e^{-\mathbf{A}x_0}\mathbf{y}_0 = e^{(x-x_0)\mathbf{A}}\mathbf{y}_0$$

An arbitrary initial condition being satisfied, formula (82) represents the general solution of equation (83).

19. Asymptotic Expansions. Asymptotic expansions introduced as early as the 18th century by H. Poincaré (1854-1912). a prominent French mathematician, are widely applied in modern mathematics. We shall consider the expansions in powers of $\frac{1}{x}$ which are more often encountered than the expansions in x. But of course this distinction is not essential because the substitution $\frac{1}{x} = x_1$ transforms an expansion in $\frac{1}{x}$ (as x approaches infinity) to an expansion in x_1 (as $x_1 \to 0$) and vice versa. For example, from series (IV.55) we directly obtain

$$e^{\frac{1}{x}} = 1 + \frac{1}{1!x} + \frac{1}{2!x^2} + \cdots + \frac{1}{n!x^n} + \cdots \tag{84}$$

Let us begin with an example. We shall investigate the behaviour of the function

$$f(x) = \int_x^\infty e^{x^2-s^2}\,ds$$

$\left[\text{which is equal to } e^{x^2}\left(\frac{\sqrt{\pi}}{2} - \mathrm{Erf}\,x\right)\right.$; see formulas (XIV.36′) and (XIV.72)$\Big]$ for $x \to \infty$. Applying L'Hospital's rule we can readily show that $f(x) \sim \frac{1}{2x}$ and hence (see Secs. III.8 and III.11) we have

$$f(x) = \frac{1}{2x} + o\left(\frac{1}{x}\right) \quad (x \to \infty) \tag{85}$$

To specify the expansion we integrate by parts:

$$\int_x^\infty e^{x^2-s^2}\,ds = \left(-e^{x^2-s^2}\frac{1}{2s}\right)\Big|_x^\infty - \frac{1}{2}\int_x^\infty \frac{e^{x^2-s^2}}{s^2}\,ds =$$

$$= \frac{1}{2x} - \frac{1}{2}\int_x^\infty \frac{e^{x^2-s^2}}{s^2}\,ds$$

We similarly verify that the last integral is equivalent to $\frac{1}{2^2x^3}$, i.e. we obtain

$$f(x) = \frac{1}{2x} - \frac{1}{2^2x^3} + o\left(\frac{1}{x^3}\right) \tag{86}$$

This as a more accurate expansion than (85) since the term $o\left(\frac{1}{x^3}\right)$ entering into (86) tends to zero faster than the analogous term in (85) as $x \to \infty$. Further integrations by parts result in still more accurate expansions

$$f(x) = \frac{1}{2x} - \frac{1}{2^2 x^3} + \frac{1 \cdot 3}{2^3 x^5} + o\left(\frac{1}{x^5}\right) \qquad (87)$$

$$f(x) = \frac{1}{2x} - \frac{1}{2^2 x^3} + \frac{1 \cdot 3}{2^3 x^5} - \frac{1 \cdot 3 \cdot 5}{2^4 x^7} + o\left(\frac{1}{x^7}\right) \qquad (88)$$

etc. (Let the reader verify the calculations!)

One can think that these operations should result in an expansion of the function $f(x)$ into the series

$$\frac{1}{2x} - \frac{1}{2^2 x^3} + \frac{1 \cdot 3}{2^3 x^5} - \frac{1 \cdot 3 \cdot 5}{2^4 x^7} + \frac{1 \cdot 3 \cdot 5 \cdot 7}{2^5 x^9} - \cdots \qquad (89)$$

but D'Alembert's test obviously indicates that this series has a zero radius of convergence, i.e. it diverges for all x! Therefore (89) cannot be used as an ordinary infinite series but formulas (85)-(88) show that we can use its partial sums.

Now we proceed to give the general definition of an asymptotic expansion. A function $f(x)$ is said to have the **asymptotic expansion**

$$f(x) \sim a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \ldots + \frac{a_n}{x^n} + \ldots \qquad (90)$$

for $x \to \infty$ if for any $n = 0, 1, 2, \ldots$ we have the representation

$$f(x) = a_0 + \frac{a_1}{x} + \ldots + \frac{a_n}{x^n} + o\left(\frac{1}{x^n}\right) \quad \text{(for } x \to \infty\text{)}$$

This property automatically holds in the case of a convergent power series of form (84). But it can also hold when series (90) is divergent everywhere or convergent to a function distinct from $f(x)$.

When applying series (90) we restrict ourselves to a certain number of terms and drop all the subsequent summands. Then estimating the last of the remaining terms we draw a conclusion as to the values of x for which the partial sum thus chosen can be used. Asymptotic expansions with alternating signs of its terms are particularly convenient because in this case the expanded function lies between any partial sum with an even number and any partial sum with an odd number.

§ 4. Trigonometric Series

29. Orthogonality. Two real functions $g(x)$ and $h(x)$ defined over a finite or infinite interval $a < x < b$ are said to be **orthogonal** to each other on the interval if

$$\int_a^b g(x)\, h(x)\, dx = 0 \qquad (91)$$

The functions are supposed to be finite or infinite but they·must have absolutely convergent integral (91). The application of the term "orthogonality" is accounted for by the fact that formula (91) turns out to be in many respects analogous to the condition of the orthogonality of two vectors given in the form of their resolutions with respect to a Cartesian basis (see Secs. VII.10 and VII.20-21).

A system of functions is referred to as being **orthogonal** on an interval if any two functions belonging to the system are orthogonal on the interval. One of the most important orthogonal systems is the system of trigonometric functions

$$1, \ \cos x, \ \sin x, \ \cos 2x, \ \sin 2x, \ \ldots, \ \cos nx, \ \sin nx, \ \ldots \quad (92)$$

which is considered on the interval $-\pi \leqslant x \leqslant \pi$. To prove the orthogonality we compute the integral

$$\int_{-\pi}^{\pi} \cos nx \cos mx \, dx = \frac{1}{2} \int_{-\pi}^{\pi} [\cos (m-n) x + \cos (m+n) x] \, dx =$$

$$= \frac{1}{2} \left[\frac{\sin (m-n) x}{m-n} + \frac{\sin (m+n) x}{m+n} \right]_{-\pi}^{\pi} = 0 \qquad (93)$$

(for $m \neq n$) and, similarly, evaluate the integrals

$$\int_{-\pi}^{\pi} \sin nx \sin mx \, dx = 0 \quad (\text{for } m \neq n)$$

and

$$\int_{-\pi}^{\pi} \cos nx \sin mx \, dx = 0 \quad (\text{for any } m, \ n = 0, 1, 2, \ldots)$$

System (92) is also orthogonal on the interval $0 \leqslant x \leqslant 2\pi$ and, generally, on each interval of length 2π. This follows from property 10 in Sec. XIV.4 if we take the product of two functions of form (92) as $f(x)$ and put $A = 2\pi$.

If we apply the property of an integral of an even function (see property 9 in Sec. XIV.4) we derive from (93) the relations

$$\int_{-\pi}^{\pi} \cos nx \cos mx \, dx = 2 \int_{0}^{\pi} \cos nx \cos mx \, dx = 0$$

$$(m, \ n = 0, 1, 2, \ldots; \ m \neq n)$$

which means that the functions

$$1, \ \cos x, \ \cos 2x, \ \ldots, \ \cos nx, \ \ldots \qquad \cdot (94)$$

form an orthogonal system on the interval $0 \leqslant x \leqslant \pi$. We similarly verify that the functions

$$\sin x, \quad \sin 2x. \ldots, \quad \sin nx, \ldots \tag{95}$$

also constitute an orthogonal system on the same interval. [Let the reader verify that system of functions (92) is not orthogonal on the interval $0 \leqslant x \leqslant \pi$.]

Changing the scale along the x-axis by introducing the scale factor $\frac{l}{\pi}$ we transform functions (92) into the functions

$$1, \quad \cos\frac{\pi x}{l}, \quad \sin\frac{\pi x}{l}, \quad \cos\frac{2\pi x}{l}, \quad \sin\frac{2\pi x}{l}, \ldots, \quad \cos\frac{n\pi x}{l},$$

$$\sin\frac{n\pi x}{l}, \ldots \tag{96}$$

which form an orthogonal system on the interval $-l \leqslant x \leqslant l$. Applying this technique we can uniformly stretch the intervals on which systems of functions (94) and (95) (and, generally, any orthogonal system) were originally defined. We can also substitute $x + h$ for x where h is an arbitrary constant. i.e. shift the graphs of the functions forming an orthogonal system along the x-axis, which does not affect the orthogonality of the system (on the shifted interval).

There are many orthogonal systems of functions other than the trigonometric functions. For instance, let us construct a system of **orthogonal polynomials** on the interval $-1 \leqslant x \leqslant 1$. Take the system

$$1, \quad x, \quad x^2, \quad x^3, \ldots, \quad x^n, \ldots \quad (-1 \leqslant x \leqslant 1) \tag{97}$$

The first two functions of the system are orthogonal to each other:

$$\int_{-1}^{1} 1 \cdot x \, dx = \frac{x^2}{2}\Big|_{-1}^{1} = 0$$

Thus we can put $P_0(x) \equiv 1$ and $P_1(x) \equiv x$. But the third function in (97) is not orthogonal to the first one (check it up!). To obtain a third function orthogonal to the former two functions let us take a linear combination of the first three functions of system (97): $P_2(x) = ax^2 + bx + c$. The coefficients a, b. c must be chosen in such a way that $P_2(x)$ be orthogonal to the polynomials $P_0(x)$ and $P_1(x)$ constructed above:

$$\int_{-1}^{1} (ax^2 + bx + c) \cdot 1 \cdot dx = 0, \quad \text{and} \quad \int_{-1}^{1} (ax^2 + bx + c) \cdot x \cdot dx = 0$$

From this we find (verify the result!):

$$b = 0, \quad a = -3c, \quad \text{i.e.} \quad P_2(x) = c(-3x^2 + 1)$$

The constant c is an arbitrary quantity here. It is usually so chosen that $P_2(1) = 1$. (Such a choice of one of the equivalent objects is called the *normalization*.) Thus we obtain $c = -\dfrac{1}{2}$, and finally we have

$$P_2(x) = \frac{3}{2} x^2 - \frac{1}{2}$$

To construct $P_3(x)$ we take a linear combination of the first four functions of system (97), i.e. $P_3(x) = ax^3 + bx^2 + cx + d$ and choose the coefficients a, b. c, d in such a way that $P_3(x)$ be orthogonal to the functions $P_0(x)$, $P_1(x)$ and $P_2(x)$ already found. Applying this condition and introducing the additional requirement $P_3(1) = 1$ we obtain, by analogy with the preceding calculations,

$$P_3(x) = \frac{5}{2} x^3 - \frac{3}{2} x$$

(let the reader verify the result!). Similarly. we find

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3), \quad P_5(x) = \frac{1}{8}(63x^5 - 70x^3 - 15x)$$

etc.

These polynomials are mutually orthogonal on the interval $-1 \leqslant x \leqslant 1$. They were investigated by Legendre in 1783-1785 and are called now the **Legendre polynomials**. The polynomials play an important role in various divisions of mathematics and physics.

This **orthogonalization** process which we have applied to system of functions (97) on the interval $-1 \leqslant x \leqslant 1$ can also be used for any system of linearly independent functions on any interval if the integrals of the squares of the functions over the interval are convergent.

21. Series in Orthogonal Functions. Let us be given a system of functions

$$g_1(x), \quad g_2(x), \ldots, \quad g_n(x), \ldots \qquad (98)$$

orthogonal on an interval $a < x < b$. We sometimes encounter the problem of expanding an arbitrary function $f(x)$ defined over the same interval into a series in functions (98), i.e. into a series of the form

$$f(x) = a_1 g_1(x) + a_2 g_2(x) + \ldots + a_n g_n(x) + \ldots = \sum_{n=1}^{\infty} a_n g_n(x) \qquad (99)$$

where a_n ($n = 1, 2, 3, \ldots$) are some numerical coefficients. This leads to the questions whether it is possible to expand any function

$f(x)$, how the coefficients a_n can be found and in what way series (99) converges.

For simplicity's sake, let us consider all the functions and the interval $a < x < b$ to be finite. The answer to the first question is dependent on the choice of system (98). If expansion (99) exists for any function $f(x)$ system of functions (98) is called **complete**. It can be proved that all the orthogonal systems mentioned in Sec. 20 are complete on the corresponding intervals.

We now proceed to determine the coefficients a_n of expansion (99) under the assumption that none of the functions (98) equals zero identically. For this purpose we multiply both sides of (99) by $g_n(x)$ and integrate the result over the interval $a \leqslant x \leqslant b$:

$$\int\limits_a^b f(x)\, g_n(x)\, dx = a_1 \int\limits_a^b g_1(x)\, g_n(x)\, dx +$$

$$+ a_2 \int\limits_a^b g_2(x)\, g_n(x)\, dx + \ldots + a_n \int\limits_a^b g_n^2(x)\, dx + \ldots$$

By the orthogonality of system (98), all the integrals on the right-hand side of the last relation are equal to zero except the integral of $g_n^2(x)$, and hence we deduce the formula for the coefficients:

$$a_n = \frac{\int\limits_a^b f(x)\, g_n(x)\, dx}{\int\limits_a^b g_n^2(x)\, dx} \qquad (n = 1, 2, 3, \ldots) \qquad (100)$$

The coefficients being uniquely defined, we conclude, in particular, that if the sum of two series of form (99) is identically equal to zero the coefficients in the same functions $g_n(x)$ in the series are also equal and that if the sum of series (99) is identically equal to zero all the coefficients are also equal to zero.

22. Fourier Series. The above general results can be applied to concrete orthogonal systems of functions. For instance, taking system (92) we conclude that any finite function defined in the interval $-\pi \leqslant x \leqslant \pi$ can be expanded in a series of the form

$$f(x) = a_1 + a_2 \cos x + a_3 \sin x + a_4 \cos 2x + a_5 \sin 2x + \ldots$$

It is convenient to change the notation of the coefficients and to put down the series as

$$f(x) = a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \ldots =$$

$$= a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \qquad (101)$$

The coefficients of the series are found by formula (100):

$$\left.\begin{array}{l} a_0 = \dfrac{\displaystyle\int_{-\pi}^{\pi} f(x)\cdot 1\,dx}{\displaystyle\int_{-\pi}^{\pi} 1^2\,dx} = \dfrac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,dx \\[4ex] a_n = \dfrac{\displaystyle\int_{-\pi}^{\pi} f(x)\cos nx\,dx}{\displaystyle\int_{-\pi}^{\pi}\cos^2 nx\,dx} = \dfrac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx \\[4ex] b_n = \dfrac{\displaystyle\int_{-\pi}^{\pi} f(x)\sin nx\,dx}{\displaystyle\int_{-\pi}^{\pi}\sin^2 nx\,dx} = \dfrac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx \quad (n\geqslant 1) \end{array}\right\} \quad (102)$$

The series with respect to systems of functions (94) or (95) are investigated in a similar way:

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos nx \quad (0\leqslant x\leqslant \pi) \tag{103}$$

$$a_0 = \frac{1}{\pi}\int_0^{\pi} f(x)\,dx, \quad a_n = \frac{2}{\pi}\int_0^{\pi} f(x)\cos nx\,dx \quad (n\geqslant 1)$$

and

$$f(x) = \sum_{n=1}^{\infty} b_n \sin nx \quad (0\leqslant x\leqslant \pi) \tag{104}$$

$$b_n = \frac{2}{\pi}\int_0^{\pi} f(x)\sin nx\,dx$$

We also often use series in functions (96) and in functions obtained from (94) or (95) by changing the scale along the x-axis:

$$f(x) = a_0 + \sum_{n=1}^{\infty}\left(a_n\cos\frac{\pi nx}{l} + b_n\sin\frac{\pi nx}{l}\right) \quad (-l\leqslant x\leqslant l) \tag{105}$$

$$a_0 = \frac{1}{2l}\int_{-l}^{l} f(x)\,dx, \quad a_n = \frac{1}{l}\int_{-l}^{l} f(x)\cos\frac{\pi nx}{l}\,dx,$$

$$b_n = \frac{1}{l}\int_{-l}^{l} f(x)\sin\frac{\pi nx}{l}\,dx \quad (n\geqslant 1),$$

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{\pi n x}{l} \quad (0 \leqslant x \leqslant l) \qquad (106)$$

$$a_0 = \frac{1}{l} \int_0^l f(x)\, dx, \quad a_n = \frac{2}{l} \int_0^l f(x) \cos \frac{\pi n x}{l}\, dx \quad (n \geqslant 1)$$

and

$$f(x) = \sum_{n=1}^{\infty} b_n \sin \frac{\pi n x}{l} \quad (0 \leqslant x \leqslant l) \qquad (107)$$

$$b_n = \frac{2}{l} \int_0^l f(x) \sin \frac{\pi n x}{l}\, dx \quad (n \geqslant 1)$$

Series (101), (103) and (104) are special cases of series (105), (106) and (107) because the former can be obtained from the latter by putting $l = \pi$. They are all called **Fourier series** after the prominent French mathematician J. Fourier (1768-1830) who for the first time applied them in his investigations in the theory of heat conductivity although such series had been used before. Formulas (102) for the **Fourier coefficients** and some other similar formulas were obtained as early as 1759 by Clairaut and in 1777 by Euler.

Consider some examples of expansions in Fourier series. Let it be necessary to expand the function $y = x$ in the interval $0 \leqslant x \leqslant l$ in series (106). For this purpose we compute the coefficients:

$$a_0 = \frac{1}{l} \int_0^l x\, dx = \frac{l}{2}$$

and

$$a_n = \frac{2}{l} \int_0^l x \cos \frac{\pi n i}{l}\, dx = \frac{2}{l}\, x\, \frac{l}{\pi n} \sin \frac{\pi n x}{l} \Big|_0^l -$$

$$- \frac{2}{l} \int_0^l \frac{l}{\pi n} \sin \frac{\pi n x}{l}\, dx = \frac{2}{\pi n}\, \frac{l}{\pi n} \cos \frac{\pi n x}{l} \Big|_0^l =$$

$$= \frac{2l}{\pi^2 n^2} (\cos \pi n - 1) = - \frac{2l}{\pi^2 n^2} [1 - (-1)^n] \quad (n \geqslant 1)$$

This yields

$$a_1 = - \frac{4l}{\pi^2 1^2}, \quad a_2 = 0, \quad a_3 = - \frac{4l}{\pi^2 3^2}, \quad a_4 = 0, \ldots$$

and finally we obtain

$$x = l \left[\frac{1}{2} - \frac{4}{\pi^2} \left(\frac{1}{1^2} \cos \frac{\pi x}{l} + \frac{1}{3^2} \cos \frac{3\pi x}{l} + \frac{1}{5^2} \cos \frac{5\pi x}{l} + \dots \right) \right] \quad (108)$$
$$(0 \leqslant x \leqslant l)$$

As another example let us take a function which is defined by means of several formulas. Namely, let

$$f(x) = \begin{cases} 1 & \text{for } -l < x < -l + \alpha \\ 0 & \text{for } -l + \alpha < x < 0 \\ 1 & \text{for } 0 < x < \alpha \\ 0 & \text{for } \alpha < x < l \end{cases}$$

where α is an arbitrary constant number belonging to the interval $0 < \alpha < l$. The graph of the function is represented in Fig. 340b
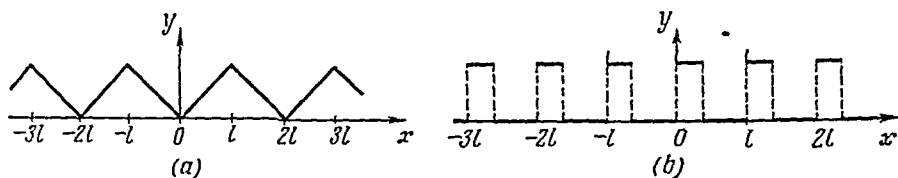
Fig. 340

by the part corresponding to the interval $-l < x < l$. Let us expand the function in series (105):

$$a_0 = \frac{1}{2l} \int_{-l}^{l} f(x)\, dx = \frac{1}{2l} \left(\int_{-l}^{-l+\alpha} f(x)\, dx + \int_{-l+\alpha}^{0} f(x)\, dx + \int_{0}^{\alpha} f(x)\, dx + \right.$$

$$\left. + \int_{\alpha}^{l} f(x)\, dx \right) = \frac{1}{2l} \left(\int_{-l}^{-l+\alpha} 1\, dx + \int_{-l+\alpha}^{0} 0\, dx + \int_{0}^{\alpha} 1\, dx + \int_{\alpha}^{l} 0\, dx \right) =$$

$$= \frac{1}{2l} (\alpha + 0 + \alpha + 0) = \frac{\alpha}{l},$$

$$a_n = \frac{1}{l} \int_{-l}^{l} f(x) \cos \frac{n\pi x}{l}\, dx = \frac{1}{l} \left(\int_{-l}^{-l+\alpha} \cos \frac{n\pi x}{l}\, dx + \right.$$

$$\left. + \int_{0}^{\alpha} \cos \frac{n\pi x}{l}\, dx \right) = \frac{1}{n\pi} \left[\sin \frac{n\pi(-l+\alpha)}{l} + \sin \frac{n\pi\alpha}{l} \right] =$$

$$= \frac{1}{n\pi} \left(-\sin \frac{n\pi l}{l} \cos \frac{n\pi\alpha}{l} + \cos \frac{n\pi l}{l} \sin \frac{n\pi\alpha}{l} + \sin \frac{n\pi\alpha}{l} \right) =$$

$$= \frac{1}{n\pi} [(-1)^n + 1] \sin \frac{n\pi\alpha}{l} \qquad (n \geqslant 1)$$

and

$$b_n = \frac{1}{l}\left(\int\limits_{-l}^{-l+\alpha} \sin\frac{n\pi x}{l}\,dx + \int\limits_{0}^{\alpha}\sin\frac{n\pi x}{l}\,dx\right) =$$

$$= -\frac{1}{n\pi}\left[\cos\frac{n\pi\,(-l+\alpha)}{l} - \cos\,(-n\pi) + \cos\frac{n\pi\alpha}{l} - 1\right] =$$

$$= -\frac{1}{n\pi}\left[\cos\frac{n\pi l}{l}\cos\frac{n\pi\alpha}{l} + \sin\frac{n\pi l}{l}\sin\frac{n\pi\alpha}{l} - (-1)^n + \cos\frac{n\pi\alpha}{l} - 1\right] =$$

$$= -\frac{1}{n\pi}\left\{[(-1)^n + 1]\cos\frac{n\pi\alpha}{l} - (-1)^n - 1\right\} =$$

$$= \frac{1}{n\pi}[(-1)^n + 1]\left(1 - \cos\frac{n\pi\alpha}{l}\right)\qquad (n \geqslant 1)$$

(Here we have applied the technique which is used for computing an integral of any function represented by several formulas.) Thus, we have $a_n = b_n = 0$ for all odd n and

$$a_{2k} = \frac{1}{k\pi}\sin\frac{2k\pi\alpha}{l},\qquad b_{2k} = \frac{1}{k\pi}\left(1 - \cos\frac{2k\pi\alpha}{l}\right)\quad (k = 1,\ 2,\ 3,\ \ldots)$$

for even numbers $n = 2k$ $(k = 1, 2, 3, \ldots)$.

The result becomes particularly simple when $\alpha = \frac{l}{2}$ because in this case

$$a_{2k} = \frac{1}{k\pi}\sin k\pi = 0,\qquad b_{2k} = \frac{1}{k\pi}(1 - \cos k\pi) =$$

$$= \frac{1}{k\pi}[1 - (-1)^k]\qquad (k = 1,\ 2,\ \ldots)$$

and thus the series takes the form

$$f(x) = \frac{1}{2} + \frac{2}{\pi}\left(\frac{1}{1}\sin\frac{2\pi x}{l} + \frac{1}{3}\sin\frac{6\pi x}{l} + \frac{1}{5}\sin\frac{10\pi x}{l} + \cdots\right)\qquad (109)$$

Consequently, a function represented by several formulas can be expanded in a unique series. The discovery of this fact by Fourier was a remarkable event which led to a considerable extension of the notion of a function.

In applying the theory to practical expansions in Fourier series we usually apply formulas of numerical integration (see Sec. XIV.13) which are particularly important when the function in question is represented by a table or a graph. For instance. suppose that we consider an expansion in series (107) and want to utilize the trapezoid rule by dividing the interval of integration into 24 parts. Then, introducing the notation $x_k = \frac{kl}{24}$, $f_k = f(x_k)$ $(k = 0, 1, \ldots, 24)$

obtain

$$b_n = \frac{2}{l} \int\limits_0^l f(x) \sin \frac{n\pi x}{l}\, dx \approx$$

$$\approx \frac{2}{l}\frac{l}{24}\left(\frac{f_0}{2}\sin\frac{n\pi x_0}{l}+f_1\sin\frac{n\pi x_1}{l}+\ldots+\frac{f_{24}}{2}\sin\frac{n\pi x_{24}}{l}\right)=$$

$$=\frac{1}{12}\left(\frac{f_0}{2}\sin 0^\circ\, n+f_1\sin 7.5^\circ\, n+f_2\sin 15^\circ\, n+\right.$$

$$\left.+\ldots+\frac{f_{24}}{2}\sin 180^\circ\, n\right) \qquad\qquad (110)$$

e see that for any n it is only the following values of the sine that
e needed:

$\sin 0^\circ\ \ = 0.0000,$	$\sin 52.5^\circ = 0.7934,$
$\sin 7.5^\circ\ = 0.1305,$	$\sin 60^\circ\ \ = 0.8660,$
$\sin 15^\circ\ \ = 0.2588,$	$\sin 67.5^\circ = 0.9239,$
$\sin 22.5^\circ = 0.3827,$	$\sin 75^\circ\ \ = 0.9659,$
$\sin 30^\circ\ \ = 0.5000,$	$\sin 82.5^\circ = 0.9914,$
$\sin 37.5^\circ = 0.6088,$	$\sin 90^\circ\ \ = 1.0000$
$\sin 45^\circ\ \ = 0.7071,$	

Vhen applying formula (110) for a given n we must substitute the
orresponding values of the sine taken from this table by applying
he reduction formulas of trigonometry, group together the terms
aving the same second factor, sum up the values of f_h in these
roups and then, after the multiplication has been performed, com-
ute the whole sum.

23. **Expanding a Periodic Function.** Fourier series are used not
only for expanding a function defined on a finite interval but also
for functions defined over the whole axis. We first suppose that
a function $f(x)$ is defined in the interval $-\pi \leqslant x \leqslant \pi$. Let us
expand it in series (101). The terms of the series are defined not
only inside the interval but also outside it and their period is equal
to 2π (see Sec. I.16) because

$$\cos n(x+2\pi) \equiv \cos(nx+2\pi n) \equiv \cos nx$$

$$\sin n(x+2\pi) \equiv \sin(nx+2\pi n) \equiv \sin nx \qquad (n=1,\, 2,\, 3,\, \ldots)$$

Hence, the sum can also be continued on the whole x-axis and
is a periodic function of period 2π. But the sum is equal to $f(x)$
on the interval $-\pi \leqslant x_i \leqslant \pi$ and consequently series (101)
extends the function $f(x)$ from the interval $-\pi \leqslant x \leqslant \pi$ onto the
whole x-axis with period 2π.

Similarly, all the terms of series (103) being even functions, its
sum results from the extension of the function $f(x)$ as an even
function from the interval $0 \leqslant x \leqslant \pi$ onto the whole x-axis with

period 2π; similarly, the sum of series (104) is the continuation of $f(x)$ as an odd function with the same period. An analogous result is obtained when we take series (105)-(107) but the period is naturally equal to $2l$.

Fig. 340 represents the graphs of the sums of the series considered in the examples of Sec. 22 which are regarded as being extended in the whole x-axis. It should be noted that although $2l$ is a period for the second example it is not the least period which equals l in this case.

Now let a function $f(x)$ be originally defined over the whole x-axis as a periodic function of period 2π. If we take a series of form (101) in which the coefficients are computed by formulas (102) then, as we have shown, its sum will yield the periodic continuation of $f(x)$ from the interval $-\pi \leqslant x \leqslant \pi$ onto the whole x-axis with period 2π and thus it will coincide with $f(x)$ on the whole axis.

The coefficients can also be found by formulas $a_0 = \dfrac{1}{2\pi} \int\limits_{\alpha-\pi}^{\alpha+\pi} f(x)\,dx,$ $a_n = \dfrac{1}{\pi} \int\limits_{\alpha-\pi}^{\alpha+\pi} f(x)\cos nx\,dx$ and $b_n = \dfrac{1}{\pi} \int\limits_{\alpha-\pi}^{\alpha+\pi} f(x)\sin nx\,dx$ $(n = 1, 2, \ldots)$ where α is an arbitrary number. In particular, we can take the formula $a_0 = \dfrac{1}{2\pi} \int\limits_{0}^{2\pi} f(x)\,dx$ and similar formulas for the coefficients a_n and b_n. This is implied by the periodicity of the integrands (see Sec. XIV.4, property 10).

Similarly, an even (odd) function with period 2π can be expanded in series (103) [series (104)]. For a function of period $2l$ we obtain series (105)-(107).

Expansion (105) is often transformed by means of formula (I.18). This results in

$$f(x) = a_0 + \sum_{n=1}^{\infty} M_n \sin\left(\frac{\pi n x}{l} + \alpha_n\right) \qquad (111)$$

The constant a_0 is equal to the mean value of the function $f(x)$ on the whole x-axis (see Sec. XIV.5) since the means of the other summands are equal to zero (check it up!). The first variable summand in (111) is called the **fundamental harmonic**; it has the least period $2l$. The subsequent summands are called **higher (upper) harmonics**. Their least periods are equal, in succession, to $\dfrac{2l}{2}, \dfrac{2l}{3}, \dfrac{2l}{4}$ and so on. Therefore an expansion of a periodic function in a Fourier series is referred to as the **harmonic analysis**.

If the independent variable is interpreted as time it is advisable to denote the period by T and to rewrite formula (111) in the form

$$f(t) = a_0 + \sum_{n=1}^{\infty} M_n \sin(n\omega t + \alpha_n) \qquad \left(\omega = \frac{2\pi}{T}\right)$$

Thus, a Fourier series represents an arbitrary periodic oscillatory motion in the form of a sum of harmonic oscillations with multiple frequencies. Such expansions are well known in acoustics where the fundamental harmonic $M_1 \sin(\omega t + \alpha_1)$ determines the **fundamental tone** and the subsequent harmonics are the **overtones** which determine the tone colour.

The periods of the summands in expansion (112) are commensurable with one another, i.e. their ratios are rational numbers. This is related to the general property that the sum of periodic functions with different periods is periodic if and only if the periods of the summands are commensurable. A sum of periodic functions with incommensurable periods belongs to a wider class of the so-called *almost periodic functions* which have many applications. In particular, they are used for investigating superpositions of non-synchronous vibrations.

24. Example. Bessel's Functions as Fourier Coefficients. A function of the form

$$e^{ix\cos t} = \cos(x\cos t) + i\sin(x\cos t) \tag{112}$$

plays an important role in radioengineering. It is an even periodic function with period 2π with respect to t for any fixed x. Therefore it can be expanded in a Fourier series of form (103). To obtain the expansion we must multiply the series

$$e^{\frac{ix}{2}}e^{it} = 1 + \frac{1}{1!}\left(\frac{ix}{2}\right)e^{it} + \frac{1}{2!}\left(\frac{ix}{2}\right)^2 e^{i2t} + \frac{1}{3!}\left(\frac{ix}{2}\right)^3 e^{i3t} + \dots$$

by the series

$$e^{\frac{ix}{2}}e^{-it} = 1 + \frac{1}{1!}\left(\frac{ix}{2}\right)e^{-it} + \frac{1}{2!}\left(\frac{ix}{2}\right)^2 e^{-i2t} + \frac{1}{3!}\left(\frac{ix}{2}\right)^3 e^{-i3t} + \dots$$

After the multiplication has been performed we obtain periodic function (112) on the left-hand side. Let us combine the terms on the right-hand side which contain the same exponential functions. To do this we note that we have the coefficient

$$1 \cdot \frac{1}{k!}\left(\frac{ix}{2}\right)^k + \frac{1}{1!}\left(\frac{ix}{2}\right)\frac{1}{(k+1)!}\left(\frac{ix}{2}\right)^{k+1} +$$

$$+ \frac{1}{2!}\left(\frac{ix}{2}\right)^2 \frac{1}{(k+2)!}\left(\frac{ix}{2}\right)^{k+2} + \dots = i^k\left(\frac{1}{0!\,k!}\left(\frac{x}{2}\right)^k - \right.$$

$$\left. - \frac{1}{1!\,(k+1)!}\left(\frac{x}{2}\right)^{k+2} + \frac{1}{2!\,(k+2)!}\left(\frac{x}{2}\right)^{k+1} - \dots\right) = i^k J_k(x)$$

in e^{ikt} and e^{-ikt} (see Sec. XV.26). Thus, we obtain

$$e^{ix \cos t} = J_0(x) + \sum_{k=1}^{\infty} i^k J_k(x)(e^{ikt} + e^{-ikt}) =$$

$$= J_0(x) + 2 \sum_{k=1}^{\infty} i^k J_k(x) \cos kt \qquad (113)$$

It is Fourier expansion (113) that we need. Separating the real and imaginary parts in (113) we arrive at the expansions

$$\cos(x \cos t) = J_0(x) - 2J_2(x) \cos 2t + 2J_4(x) \cos 4t - \ldots$$

and

$$\sin(x \cos t) = 2J_1(x) \cos t - 2J_3(x) \cos 3t + 2J_5(x) \cos 5t - \ldots$$

Formula (113) implies, in particular, *the integral representation of a Bessel function* of an integral order:

$$J_n(x) = \frac{1}{\pi i^n} \int_0^{\pi} e^{ix \cos t} \cos nt \, dt$$

[check up the result taking advantage of formula (103) for the coefficients of series (103)].

25. Speed of Convergence of a Fourier Series. Let a bounded periodic function $f(x)$ of period $2l$ be expanded in Fourier series (105). We can easily show that all the Fourier coefficients are bounded above in their absolute values by the same positive constant:

$$|a_n| = \frac{1}{l} \left| \int_{-l}^{l} f(x) \cos \frac{n\pi x}{l} \, dx \right| \leqslant$$

$$\leqslant \frac{1}{l} \int_{-l}^{l} |f(x)| \left| \cos \frac{n\pi x}{l} \right| dx \leqslant \frac{1}{l} \int_{-l}^{l} |f(x)| \, dx$$

The same result is obtained if the function $f(x)$ is unbounded but absolutely integrable (summable) on the interval $-l \leqslant x \leqslant l$, i.e. if

$$\int_{-l}^{l} |f(x)| \, dx < \infty$$

Let the function $f(x)$ be bounded and discontinuous. Then its Fourier coefficients a_n and b_n are of the order of $\frac{1}{n}$ as $n \to \infty$ (in

particular, this is the case in the second example considered in Sec. 22).

Actually, suppose, for definiteness, that $f(x)$ has two discontinuities in the interval $-l \leqslant x \leqslant l$ at the points $x = x_1$ and $x = x_2$ and that $-l < x_1 < x_2 < l$. Then we have

$$a_n = \frac{1}{l} \int_{-l}^{l} f(x) \cos \frac{n\pi x}{l} \, dx = \frac{1}{l} \int_{-l}^{x_1} f(x) \cos \frac{n\pi x}{l} \, dx +$$

$$+ \frac{1}{l} \int_{x_1}^{x_2} f(x) \cos \frac{n\pi x}{l} \, dx + \frac{1}{l} \int_{x_2}^{l} f(x) \cos \frac{n\pi x}{l} \, dx$$

Let us integrate by parts each of these integrals:

$$a_n = \frac{1}{n\pi} \left[f(x_1 - 0) \sin \frac{n\pi x_1}{l} - f(-l) \sin \frac{n\pi(-l)}{l} - \right.$$

$$\left. - \int_{-l}^{x_1} f'(x) \sin \frac{n\pi x}{l} \, dx \right] + \ldots = -\frac{1}{n\pi} [f(x_1 + 0) -$$

$$- f(x_1 - 0)] \sin \frac{n\pi x_1}{l} - \frac{1}{n\pi} [f(x_2 + 0) -$$

$$- f(x_2 - 0)] \sin \frac{n\pi x_2}{l} - \frac{1}{n\pi} \int_{-l}^{l} f'(x) \sin \frac{n\pi x}{l} \, dx \qquad (114)$$

The last summand on the right-hand side differs from the Fourier coefficient of $f'(x)$ only in the constant factor in front of the integral.

But $\int_{-l}^{l} |f'(x)| \, dx < \infty$ because the derivative $f'(x)$ retains its sign on each interval (α, β) of monotonicity and continuity of the function $f(x)$ and hence

$$\int_{\alpha}^{\beta} |f'(x)| \, dx = \left| \int_{\alpha}^{\beta} f'(x) \, dx \right| = |f(\beta - 0) - f(\alpha + 0)| < \infty$$

We have excluded from our considerations functions having an infinite number of intervals of monotonicity within a finite interval of variation of x (see Fig. 102) because they are rarely encountered. We see that under the assumptions concerning $f(x)$ the last integral in formula (114) is bounded which implies what has been said about the order of smallness of the Fourier coefficients.

Now let the function $f(x)$ itself be continuous. Suppose that its derivative has discontinuities and is bounded (this is the case in the first example considered in Sec. 22). Then its Fourier coefficients are of the order of $\frac{1}{n^2}$ as $n \to \infty$.

Indeed, taking the expression of a Fourier coefficient and integrating by parts we obtain

$$a_n = \frac{1}{l} \int_{-l}^{l} f(x) \cos \frac{n\pi x}{l}\, dx = -\frac{1}{n\pi} \int_{-l}^{l} f'(x) \sin \frac{n\pi x}{l}\, dx$$

Applying the argument of the preceding paragraphs to the integral on the right-hand side we conclude that it is of the order of $\frac{1}{n}$ as $n \to \infty$, and hence a_n is of the order of $\frac{1}{n^2}$ (the same result is similarly obtained for b_n). If $f(x)$ and $f'(x)$ are continuous and $f''(x)$ has discontinuities we can perform the integration by parts twice and thus prove that in this case the Fourier coefficients are of the order of $\frac{1}{n^3}$ and so on. Hence, the order of smallness of Fourier coefficients depends on the "smoothness" of the function in question, i.e. on the number of continuous derivatives it possesses. The greater the number, the higher the order of smallness of the Fourier coefficients, that is the higher the speed of convergence of the Fourier series of the function.
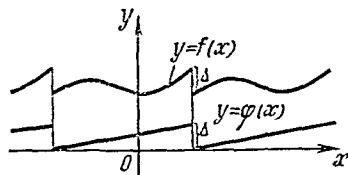
Fig. 341

The Fourier series of a discontinuous function $f(x)$ converges very slowly and therefore it is difficult to apply it to practical calculations. To overcome the difficulty we sometimes try to construct a function $\varphi(x)$ having discontinuities at the same points as $f(x)$ and the same jumps (see Fig. 341). It is advisable to choose a function $\varphi(x)$ whose structure is as simple as possible. After $\varphi(x)$ has been chosen we form the difference $f(x) - \varphi(x)$ which no longer has discontinuities and therefore is expanded in a Fourier series whose speed of convergence is higher than that of the Fourier series of $f(x)$. Consequently, we represent $f(x)$ as the sum of a simple function $\varphi(x)$ and a Fourier series with a higher speed of convergence. We can similarly eliminate the discontinuities of the first derivative and so on (compare this with the methods used in Sec. 4).

Thus, if a function $f(x)$ is continuous and has a bounded derivative the order of smallness of its Fourier coefficients (as $n \to \infty$) is not less than that of $\frac{1}{n^2}$. But according to Sec. 1

$$\sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

and hence, by Weierstrass' test (see Sec. 8), the Fourier series uniformly converges on the whole x-axis. A more extensive investigation

shows that the condition which we have imposed on $f'(x)$ is unnecessary because Weierstrass' test can be applied under some more general assumptions. In the case of a continuous function the substitution of any numerical value of x into the series exactly yields the corresponding value $f(x)$.

If a function $f(x)$ is discontinuous its Fourier series by no means converges uniformly because its terms are continuous functions (see property 1 in Sec. 9). It can be shown that in this case the substitution of a numerical value of x into the series results in $f(x)$ at all the points of continuity of $f(x)$ and in

$$\frac{f(x-0)+f(x+0)}{2}$$

at the points of discontinuity. For instance, series (109) has the sum equal to $\frac{1}{2}$ for $x = 0$ since in this case $f(-0) = 0$ and $f(+0) = 1$.

If a function defined in a finite interval is expanded in a Fourier series (see Sec. 22) the speed of convergence of the series is specified by the discontinuities of the function and of its derivatives which occur after the function has been periodically extended onto the whole x-axis as it was described in Sec. 23. For instance, the Fourier coefficients in the first example considered in Sec. 22 are of the order of $\frac{1}{n^2}$ since the extension (see Fig. 340a) results in a function whose first derivative has discontinuities. If the same function is expanded into a series in sines [series (107)] the coefficients are of the order of $\frac{1}{n}$ because after the extension has been performed the function itself has discontinuities (why?).

Fourier series enable us to find the sums of many interesting numerical series. For instance, if we substitute $x = \frac{l}{4}$ into series (109) we obtain

$$1 = \frac{1}{2} + \frac{2}{\pi}\left(\frac{1}{1}\sin\frac{\pi}{2} + \frac{1}{3}\sin\frac{3\pi}{2} + \frac{1}{5}\sin\frac{5\pi}{2} + \ldots\right) =$$
$$= \frac{1}{2} + \frac{2}{\pi}\left(\frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \ldots\right)$$

which implies

$$\frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots = \frac{\pi}{4}$$

If we substitute $x = 0$ into series (108) we get

$$0 = l\left[\frac{1}{2} - \frac{4}{\pi^2}\left(\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \ldots\right)\right]$$

i.e.

$$\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \ldots = \frac{\pi^2}{8}$$

Further, the last result makes it possible to find the value $\zeta(2)$ of the zeta function [see formula (19)] :

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \ldots = \frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \ldots +$$

$$+ \frac{1}{2^2}\left(\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots\right) = \frac{\pi^2}{8} + \frac{\zeta(2)}{4}$$

which results in

$$\zeta(2) = \frac{\pi^2}{6} = 1.645$$

26. Fourier Series in Complex Form. Applying Euler's formulas we can pass from a Fourier series containing trigonometric functions to a series in exponential functions which is sometimes preferable. To perform such a transformation of series (105) we can take the formulas

$$\cos\frac{\pi n x}{l} = \frac{e^{\frac{i\pi n x}{l}} + e^{-\frac{i\pi n x}{l}}}{2}$$

and

$$\sin\frac{\pi n x}{l} = \frac{e^{\frac{i\pi n x}{l}} - e^{-\frac{i\pi n x}{l}}}{2i}$$

After the substitution is performed we combine similar terms on the right-hand side and thus obtain a series in which the summation is extended over all the exponential functions of the form $e^{\frac{i\pi n x}{l}}$ ($n = 0, \pm 1, \pm 2, \ldots$). Thus we obtain

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{\frac{i\pi n x}{l}} \qquad (115)$$

where c_n ($n = 0, \pm 1, \pm 2, \ldots$) are some complex coefficients.*

To find the coefficients c_n we multiply both sides by $e^{-\frac{i\pi n x}{l}}$, for a fixed n, and integrate the result from $-l$ to l. The integrals

---

* A series of this type in which the summation index runs from $-\infty$ to ∞ is sometimes referred to as a *two-way series*. — *Tr.*

of the terms with numbers different from n are equal to zero:

$$\int_{-l}^{l} c_m e^{\frac{i\pi mx}{l}} e^{-\frac{i\pi nx}{l}} dx = c_m \frac{e^{\frac{i\pi (m-n)}{l}}}{i\pi (m-n)} l \Big|_{-l}^{l} =$$

$$= \frac{c_m l}{i\pi (m-n)} [e^{i\pi (m-n)} - e^{-i\pi (m-n)}] =$$

$$= \frac{2c_m l}{\pi (m-n)} \sin \pi (m-n) = 0 \quad (m \neq n)$$

The integration of the term having the number n results in $2lc_n$. Thus, the coefficients in series (115) are computed by the formula

$$c_n = \frac{1}{2l} \int_{-l}^{l} f(x) e^{-\frac{i\pi nx}{l}} dx \quad (n = 1, 2, \ldots) \tag{116}$$

Let us discuss the general case of an expansion with respect to an orthogonal system of complex functions. Two complex functions $g(x)$ and $h(x)$ are said to be *orthogonal* to each other on the interval $a \leqslant x \leqslant b$ if

$$\int_{a}^{b} g(x) h^*(x) dx = 0 \tag{117}$$

where the asterisk designates the complex conjugate function (see Sec. VIII.3). In a special case when g and h are real this definition coincides with former definition (91). It should be noted that if we pass from (117) to the complex conjugate expression in both sides we obtain

$$\left(\int_{a}^{b} g(x) h^*(x) dx \right)^* = \int_{a}^{b} (g(x) h^*(x))^* dx = \int_{a}^{b} h(x) g^*(x) dx = 0$$

which means that the orthogonality condition is independent of the order in which we enumerate the functions, that is if g is orthogonal to h it follows that h is orthogonal to g, and hence we can speak about the mutual orthogonality of the functions.

If we have a system of complex functions

$$g_1(x), \quad g_2(x), \ldots, \quad g_n(x), \ldots \tag{118}$$

which are orthogonal on an interval $a \leqslant x \leqslant b$ and if a complex function $f(x)$ can be expanded into a series in these functions [this is always the case when system (108) is complete] we obtain

$$f(x) = c_1 g_1(x) + c_2 g_2(x) + \ldots + c_n g_n(x) + \ldots = \sum_{n=1}^{\infty} c_n g_n(x) \tag{119}$$

To determine the coefficients we multiply both sides by $g_n^*(x)$ for a fixed n and integrate the result from a to b. By the orthogonality condition, only one term on the right-hand side will be different from zero and hence we obtain the formula

$$c_n = \frac{\int\limits_a^b f(x)\, g_n^*(x)\, dx}{\int\limits_a^b g_n(x)\, g_n^*(x)\, dx} = \frac{\int\limits_a^b f(x)\, g_n^*(x)\, dx}{\int\limits_a^b |g_n(x)|^2\, dx}$$

Expansion (115) is a particular case of (119) when the system of functions

$$\ldots, e^{-\frac{i2\pi x}{l}},\quad e^{-\frac{i\pi x}{l}},\quad 1,\quad e^{\frac{i\pi x}{l}},\quad e^{\frac{i2\pi x}{l}},\ \ldots$$

(which is complete on the interval $-l \leqslant x \leqslant l$) is taken as system (118). The expansion is valid for any bounded function and even for an unbounded complex function $f(x)$ which is absolutely integrable over the interval (see Sec. XIV.16).

27. Parseval Relation. Let us return to real functions. Take an orthogonal and complete system of functions on an interval $a \leqslant \leqslant x \leqslant b$:

$$g_1(x),\quad g_2(x),\ \ldots,\quad g_n(x),\ \ldots \tag{120}$$

Let us consider the expansion of an arbitrary function $f(x)$ into a series in these functions. Squaring both sides of the expansion and integrating the result from a to b we arrive at the integral

$$\int\limits_a^b f^2(x)\, dx$$

on the left-hand side. We shall suppose that the integral has a finite numerical value. After the right-hand side is squared we obtain the sum of the squares of the terms and the products of different terms taken pairwise. The integrals of the latter are equal to zero according to the orthogonality of functions (120) whereas the integrals of the former are different from zero, and thus we have

$$\int\limits_a^b f^2(x)\, dx = \sum_{n=1}^{\infty} a_n^2 \int\limits_a^b g_n^2(x)\, dx \tag{121}$$

This formula is referred to as the **Parseval relation** (Parseval theorem). In the particular case of a Fourier series of form (101)

we obtain

$$\int\limits_{-\pi}^{\pi} f^2(x)\,dx = 2\pi a_0^2 + \pi \sum_{n=1}^{\infty} (a_n^2 + b_n^2)$$

and for series (105) we get

$$\int\limits_{-l}^{l} f^2(x)\,dx = 2la_0^2 + l \sum_{n=1}^{\infty} (a_n^2 + b_n^2)$$

The relations were found in 1805. By the way, they directly imply that $a_n \to 0$ and $b_n \to 0$ as $n \to \infty$.

If a system of type (120) is incomplete it is possible to prove that it can be completed. After the completion, relation (121) becomes true. But all the terms on the right-hand side are non-negative and hence if a system of orthogonal functions is not complete we have the inequality

$$\sum_{n=1}^{\infty} a_n^2 \int\limits_{a}^{b} g_n^2(x)\,dx \leqslant \int\limits_{a}^{b} f^2(x)\,dx$$

for any function $f(x)$. The sign of equality occurs here only for those functions $f(x)$ which can be expanded with respect to system of functions (120).*

Parseval's equality (121) enables us to apply a new approach to constructing a series in orthogonal functions. Let us be given a finite number of functions

$$g_1(x), \quad g_2(x), \ldots, \quad g_n(x) \qquad ' \qquad (122)$$

which are orthogonal on an interval $a \leqslant x \leqslant b$. We now pose the following problem: it is required to form a linear combination of functions (122) whose mean square deviation from a given function $f(x)$ (see Sec. 7) is minimal.

To solve the problem let us consider functions (122) to be a part of a complete orthogonal system of form (120). Then

$$f(x) - \sum_{k=1}^{n} C_k g_k(x) = \sum_{k=1}^{\infty} a_k g_k(x) - \sum_{k=1}^{n} C_k g_k(x) =$$

$$= \sum_{k=1}^{n} (a_k - C_k) g_k(x) + \sum_{k=n+1}^{\infty} a_k g_k(x)$$

where a_k are the Fourier coefficients of the expansion of $f(x)$ with respect to functions (120) and C_k are arbitrary constants. By equality

---

* The last relation is referred to as Bessel's inequality. — *Tr.*

(121), we can write

$$\int_a^b \left[f(x) - \sum_{h=1}^n C_h g_h(x) \right]^2 dx = \sum_{h=1}^n (a_h - C_h)^2 \int_a^b g_h^2(x)\,dx +$$
$$+ \sum_{h=n+1}^\infty a_h^2 \int_a^b g_h^2(x)\,dx$$

The last formula indicates that if the coefficients C_1, C_2, ..., C_n are varied in an arbitrary way the minimal value of the right-hand side is attained when

$$C_1 = a_1, \quad C_2 = a_2, \ldots, \quad C_n = a_n$$

Thus, to obtain a linear combination of a fixed number of orthogonal functions (122) whose mean square deviation from $f(x)$ is minimal we must take the corresponding partial sum of the expansion of $f(x)$ into a series with respect to the system of orthogonal functions (120), that is the linear combinations with the coefficients defined by formulas (100).

An analogous argument applied to a complete orthogonal system of complex functions of type (118) and to series (119) leads to the formula

$$\int_a^b |f(x)|^2\,dx = \sum_{n=1}^\infty |a_n|^2 \int_a^b |g_n(x)|^2\,dx \tag{123}$$

if we take into account the equality $aa^* = |a|^2$.

28. Hilbert Space. We now return to real functions defined on a finite interval $a \leqslant x \leqslant b$. It turns out that there exists a far-reaching analogy between such functions and vectors which we mentioned in Sec. 20. Since we can perform linear operations on the functions according to ordinary algebraic rules the functions form a linear space in the sense of the definition given in Sec. VII.17. Moreover, if we introduce the notion of a scalar product of two functions by means of the formula

$$(f, g) = \int_a^b f(x)\,g(x)\,dx \tag{124}$$

we can readily verify that all the axioms enumerated in Sec. VII.20 are fulfilled and thus we obtain a space which is not only linear but Euclidean as well. We include in this space all the bounded functions and also all the unbounded functions with *integrable (summable)*

res, that is such functions $f(x)$ that

$$(f, f) = \int_a^b f^2(x) \, dx < \infty \tag{125}$$

mple inequality $2 \mid f(x) \, g(x) \mid \leqslant f^2(x) + g^2(x)$ implies that
il (124) of a function satisfying condition (125) is convergent
led it is an improper integral).
set of functions satisfying condition (125) equipped with
product (124) is called the Hilbert space L_2 after the famous
n mathematician D. Hilbert (1862-1943). This is an infinite-
sional Euclidean space (see Sec. VII.20). By definition (124),
ion (91) is nothing but the orthogonality condition for vectors
ing to this space. A complete orthogonal system of functions
rthogonal basis in the space L_2. It should be noted that when
ng the notion of a complete system of functions (see Sec. 21)
space L_2 we must interpret the convergence of series (99)
convergence in the mean square (see Sec. 8). Thus, the con-
ice in L_2 is the convergence in the mean square. Formulas (100)
coefficients of an expansion are a particular case of formulas
9) and the orthogonalization process described in Sec. 20
ling but a realization of the process discussed in Sec. VII.21.
r, Parseval's equality (121) in the Hilbert space L_2 is ana-
to Pythagoras' theorem: the square of the diagonal of a rec-
ar parallelepiped is equal to the sum of squares of all its
sions. (We suggest that the reader try to interpret geometri-
the property of partial sums of series (99) which, as it was
l in Sec. 27, minimize the mean square deviation.)
haracteristic feature of a Hilbert space is that it is infinite-
sional. This property makes it difficult to test the completeness
orthogonal system of functions. A system of k pairwise ortho-
nonzero vectors belonging to an n-dimensional Euclidean space
iplete if $k = n$ and incomplete if $k < n$. In contrast to it
inite orthogonal system containing infinitely many functions
ling to an infinite-dimensional space may not be complete.
the number of functions does not enable us to answer the
on whether a given orthogonal system is complete in the
f an infinite-dimensional space. This is a rather difficult pro-
and we shall not consider it here.
an be proved that any incomplete orthogonal system is a part
omplete orthogonal system. Therefore a function can be expan-
rith respect to the former if and only if its expansion with
t to the latter has zero coefficients in those functions of the
n which are not contained in the former.
Sec. VII.20 we proved that for Euclidean spaces there is an
tant inequality of form (VII.26). Applying it to a functional

space and squaring we deduce the inequality

$$\left(\int_a^b f(x)\,g(x)\,dx\right)^2 \leqslant \left(\int_a^b f^2(x)\,dx\right)\left(\int_a^b g^2(x)\,dx\right)$$

Substituting $|f|$ for f and 1 for g into the inequality we get

$$\left(\int_a^b |f(x)|\,dx\right)^2 \leqslant (b-a)\int_a^b f^2(x)\,dx$$

This implies that all the functions belonging to L_2 are summab
and that convergence in the mean square implies convergenc
in the mean*. The converse may not be true. By the way,
is possible to prove that the Fourier series of any summable fun
tion converges to the function in the mean.

29. **Orthogonality with Weight Function.** When we integra
a function $f(x)$ over an interval $a \leqslant x \leqslant b$ all the values of x b
longing to the interval are equivalent. But if we want to stre:
the importance of a certain value of x in comparison with the othe
we introduce a weight function (weighting function) $\rho(x) \geqslant 0$ whe
performing the integration:

$$\int_a^b f(x)\,\rho(x)\,dx$$

The function $\rho(x)$ is so chosen that its values should be greater f(
the values of x which are considered to be more important.

Two functions $g(x)$ and $h(x)$ are said to be **orthogonal with weig**
function $\rho(x)$ on an interval $(a,\ b)$ if

$$\int_a^b f(x)\,g(x)\,\rho(x)\,dx = 0$$

The whole theory of series in orthogonal functions presented i
Secs. 20, 21, 26 and 27 is directly extended to functions orthogon:
with weight function; to do this we must simply introduce th
factor $\rho(x)$ under the sign of integration in all the formulas.

An integral involving a weighting function $\rho(x)$ can be easil
transformed to an integral without a weighting function (i.e. wit
the weighting function $\rho \equiv 1$) by means of change of independer

---

* Convergence in the mean is understood here as convergence in tl
mean of order one; see footnote on page 663. — *Tr.*

variable:

$$\rho(x)\,dx = d\bar{x}, \quad \bar{x} = \int_{x_0}^{x} \rho(x)\,dx = \bar{x}(x), \quad \bar{x}(a) = \bar{a} \quad \text{and} \quad \bar{x}(b) = \bar{b}$$

(126)

Indeed, if we introduce the notation $f(x) = f(x(\bar{x})) = \bar{f}(\bar{x})$ we obtain

$$\int_{a}^{b} f(x)\,\rho(x)\,dx = \int_{\bar{a}}^{\bar{b}} \bar{f}(\bar{x})\,d\bar{x}$$

Under such a transformation any system of functions orthogonal with weight function ρ turns into a system of functions orthogonal in the sense of our former definition (Sec. 20). But nevertheless it is sometimes convenient to consider functions orthogonal with weight function without transforming them according to formula (126).

Conversely, change of variable (126) enables us to pass from any orthogonal (in the ordinary sense) system of functions to a system orthogonal with weight function ρ. Such a transformation yields the corresponding transformation of the expansions in series and therefore a complete system is transformed to a-complete one. For instance, taking the complete orthogonal system of functions

$$1, \quad \cos \bar{x}, \quad \cos 2\bar{x}, \ \ldots, \quad \cos n\bar{x}, \ \ldots \quad (0 \leqslant \bar{x} \leqslant \pi)$$

[see system (94)] and performing the transformation

$$\bar{x} = \pi - \arccos x, \quad d\bar{x} = \frac{1}{\sqrt{1-x^2}}\,dx \quad (-1 \leqslant x \leqslant 1)$$

we obtain the functions

$$T_0(x) = 1, \ T_1(x) = \cos \arccos x = x,$$
$$T_2(x) = \cos(2 \arccos x) = 2x^2 - 1,$$
$$T_3(x) = \cos(3 \arccos x) = 4x^3 - 3x, \ \ldots,$$
$$T_n(x) = \cos(n \arccos x), \ \ldots$$

after an inessential change of the signs has been made (check it up!).

These polynomials were introduced by Chebyshev in 1857. They are referred to as **Chebyshev's polynomials.** We see that they form a complete system of functions orthogonal with weight function $\frac{1}{\sqrt{1-x^2}}$ on the interval $-1 \leqslant x \leqslant 1$. The polynomials can also be obtained from system of functions (97) by means of the orthogonalization process with this weight function (let the reader perform the transformation!).

To eliminate a weight function we can also apply the following method: if we are given a system of functions

$$g_1(x), \; g_2(x), \; \ldots, \; g_n(x), \; \ldots \tag{127}$$

orthogonal with weight function $\rho(x)$ on an interval $a < x < b$ the system of functions

$$g_1(x)\sqrt{\rho(x)}, \quad g_2(x)\sqrt{\rho(x)}, \; \ldots, \quad g_n(x)\sqrt{\rho(x)}, \; \ldots \tag{128}$$

is orthogonal without weight function on the same interval (check it up!). To expand a function $f(x)$ in a series with respect to system (127) it is sufficient to expand the function $f(x)\sqrt{\rho(x)}$ with respect to system (128) and then cancel out the factor $\sqrt{\rho(x)}$.

30. Multiple Fourier Series. When expanding a function of several independent variables in a series we usually take a system of functions dependent on several indices whose number equals the number of the arguments. Then we arrive at multiple series as in Sec. 17.

The theory of multiple series with respect to systems of orthogonal functions is developed by analogy with the theory of ordinary series. For the sake of simplicity, we restrict ourselves to the case of functions of two arguments. In this case functions forming a complete orthogonal system in a domain D must depend on two indices, that is have the form $\varphi_{mn}(x, y)$ where m and n assume some discrete values. A series with respect to such a system is of the form

$$f(x, y) = \sum_{m,\, n} a_{mn}\varphi_{mn}(x, y)$$

where the symbol $\sum\limits_{m,n}$ denotes the corresponding two-fold sum. The coefficients are found after a manner of Sec. 21:

$$a_{mn} = \frac{\displaystyle\iint_D f(x, y)\, \varphi_{mn}(x, y)\, dx\, dy}{\displaystyle\iint_D \varphi_{mn}^2(x, y)\, dx\, dy} \tag{129}$$

There is a method of constructing a system of orthogonal functions of several arguments on the basis of some given systems of functions of one argument. Let us be given two complete orthogonal systems of functions of one independent variable:

$$g_1(x), \quad g_2(x), \; \ldots, \quad g_n(x), \; \ldots \quad (a \leqslant x \leqslant b)$$

and

$$h_1(x), \quad h_2(x), \; \ldots, \quad h_n(x), \; \ldots \quad (c \leqslant x \leqslant d)$$

Then the system of functions

$$\varphi_{mn}(x, y) = g_m(x)\, h_n(y) \quad (m, \; n = 1, \; 2, \; \ldots) \tag{130}$$

is orthogonal and complete in the rectangle Π: $a \leqslant x \leqslant b$, $c \leqslant$ $\leqslant y \leqslant d$.

The orthogonality of the system is implied by the relation

$$\iint\limits_{\Pi} \varphi_{mn}(x, y)\, \varphi_{\overline{mn}}(x, y)\, dx\, dy = \int\limits_a^b dx \int\limits_c^d g_m(x)\, h_n(y)\, g_{\overline{m}}(x)\, h_{\overline{n}}(y)\, dy =$$

$$= \left(\int\limits_a^b g_m(x)\, g_{\overline{m}}(x)\, dx \right) \left(\int\limits_c^d h_n(y)\, h_{\overline{n}}(y)\, dy \right)$$

which always yields zero except the case when we simultaneously have $m = \overline{m}$ and $n = \overline{n}$. The completeness can also be easily proved. Namely, given an arbitrary function $f(x, y)$ defined in Π, we can expand it with respect to functions $h_n(y)$ for any fixed x:

$$f(x, y) = \sum_{n=1}^{\infty} A_n(x)\, h_n(y)$$

where the coefficients of the expansion are dependent on x. Now we can expand these coefficients $A_n(x)$ with respect to the functions $g_m(x)$ which results in a double series with respect to system of functions (130):

$$f(x, y) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{mn} g_m(x)\, h_n(y)$$

Thus we have obtained what we set out to prove. Hence the expansion is possible.

Taking two systems of functions of form (95) defined on extended intervals $0 \leqslant x \leqslant l_1$ and $0 \leqslant y \leqslant l_2$ we can form a complete orthogonal system of functions on the corresponding rectangle:

$$\varphi_{mn}(x, y) = \sin\frac{m\pi x}{l_1} \sin\frac{n\pi y}{l_2} \qquad (m, n = 1, 2, \ldots)$$

An expansion with respect to this system has the form

$$f(x, y) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{mn} \sin\frac{m\pi x}{l_1} \sin\frac{n\pi y}{l_2} \quad (0 \leqslant x \leqslant l_1,\ 0 \leqslant y \leqslant l_2)$$

The coefficients in the series are found on the basis of formula (129):

$$a_{mn} = \frac{4}{l_1 l_2} \int\limits_0^{l_1} dx \int\limits_0^{l_2} f(x, y) \sin\frac{m\pi x}{l_1} \sin\frac{n\pi y}{l_2}\, dy$$

(verify the result!).

31. Application to the Equation of Oscillations of a String. Fourier series have many applications in the theory of equations of mathe-

matical physics. Here we shall give an example of such an application to the problem of solving the equation of small free transverse oscillations of a taut string.*

We shall consider the string to be weightless. Suppose the string has a finite length l. Let us draw the x-axis along the string in its equilibrium state so that the ends of the string have the coordinates $x = 0$ and $x = l$, respectively. We shall study plane oscillations and denote by $u\,(x,\ t)$ the transverse deflection of the point of the string with abscissa x at the moment t from the equilibrium state. In the theory of equations of mathematical physics it is proved that the function $u\,(x,\ t)$ satisfies the following partial differential equation of the second order:

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \tag{131}$$

Here a is a constant ($a = \sqrt{\dfrac{T}{\rho}}$ where T is the tension of the string and $\rho = $ const is its linear density). We shall consider the ends of the string to be fixed. Then the corresponding boundary conditions expressing this fact can be written in the form

$$u\,|_{x=0} = 0, \quad u\,|_{x=l} = 0 \quad \text{(for all } t) \tag{132}$$

We shall also suppose that the deflections and velocities of the points of the string at the initial moment of time $t = 0$ are known. Then we can write the initial conditions of the form

$$u\,|_{t=0} = \varphi\,(x), \qquad \frac{\partial u}{\partial t}\bigg|_{t=0} = \psi\,(x) \tag{133}$$

where $\varphi\,(x)$ and $\psi\,(x)$ are some given functions. Hence, we arrive at the following mathematical problem: it is required to solve equation (131) for boundary conditions (132) and initial conditions (133).

To solve the problem let us look for the sought-for solution in the form of a series of type (107) for each fixed $t \geqslant 0$. Then the coefficients will be dependent on t and thus we get

$$u\,(x,\ t) = \sum_{n=1}^{\infty} b_n\,(t) \sin \frac{\pi n x}{l} \tag{134}$$

To find the coefficients $b_n\,(t)$ we substitute this expression into equation (131). This results in

$$\sum_{n=1}^{\infty} b_n''\,(t) \sin \frac{\pi n x}{l} = -a^2 \sum_{n=1}^{\infty} b_n\,(t) \frac{\pi^2 n^2}{l^2} \sin \frac{\pi n x}{l}$$

---

* Equations of mathematical physics (and, in particular, the application of Fourier series to solving the equation of oscillation of a string) are treated in greater detail in the Appendix at the end of this English edition.—*Tr.*

that is

$$b_n^{"}(t) = -\frac{a^2\pi^2 n^2}{l^2} b_n{}'(t)$$

Solving this ordinary linear differential equation with constant coefficients by applying the methods given in Sec. XV.17 we derive

$$b_n(t) = A_n \cos\frac{a\pi n}{l} t + B_n \sin\frac{a\pi n}{l} t$$

By (134), it follows that

$$u = \sum_{n=1}^{\infty} \left(A_n \cos\frac{a\pi n}{l} t + B_n \sin\frac{a\pi n}{l} t \right) \sin\frac{\pi n}{l} x \qquad (135)$$

To determine the coefficients A_n and B_n $(n = 1, 2, 3, \ldots)$ we take advantage of initial conditions (133). This yields

$$\varphi(t) = \sum_{n=1}^{\infty} A_n \sin\frac{\pi n}{l} x, \qquad \psi(t) = \sum_{n=1}^{\infty} B_n \frac{a\pi n}{l} \sin\frac{\pi n}{l} x \qquad (0 \leqslant x \leqslant l)$$

(check up the result!). We have arrived at Fourier expansions of form (107) from which we find

$$A_n = \frac{2}{l} \int_0^l \varphi(t) \sin\frac{\pi n x}{l} dx \quad \text{and} \quad B_n = \frac{2}{a\pi n} \int_0^l \psi(t) \sin\frac{\pi n x}{l} dx$$

Substituting these quantities into (135) we thus obtain the sought-for solution. The boundary conditions are automatically satisfied here because of the properties of the functions $\sin\frac{\pi n x}{l}$ $(n = 1, 2, 3, \ldots)$ which satisfy conditions (132).

§ 5. Fourier Transformation

32. Fourier Transform. Let us take formula (115) which represents any finite function $f(x)$ defined over an interval $-l < x < l$ in the form of a complex Fourier series. This representation involves *complex harmonics.* i.e. the functions e^{ikx} with *wave numbers* $k = k_n$ where

$$k_n = \frac{\pi n}{l} \qquad (n = \ldots, -2, -1, 0, 1, 2, \ldots) \qquad (136)$$

The set of these numbers (it is depicted in Fig. 342) is called the spectrum of wave numbers. It is discrete, that is it consists of separate points to each of which there corresponds a harmonic $e^{ik_n x}$ in expansion (115) with *complex amplitude* c_n. The quantity c_n $(n = 0. \pm 1.$

± 2, ...) is defined by the formula

$$c_n = \frac{1}{2l} \int\limits_{-l}^{l} f(x)\, e^{-ik_n x}\, dx \qquad (137)$$

which is implied by formula (116).

For the sake of simplicity, we first suppose that the function $f(x)$ identically vanishes outside a finite interval $a \leqslant x \leqslant b$.
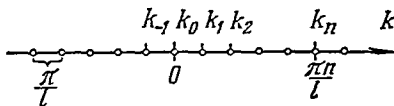
Fig. 342

Introduce the notation

$$\hat{f}(k) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f(x)\, e^{-ikx}\, dx \qquad (138)$$

This integral is in fact taken only over the interval $a \leqslant x \leqslant b$. If l is sufficiently large we can rewrite formula (137) of a Fourier coefficient in the form

$$c_n = \frac{\pi}{l}\, \frac{1}{2\pi} \int\limits_{a}^{b} f(x)\, e^{-ik_n x}\, dx = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f(x)\, e^{-ik_n x}\, dx\, \frac{\pi}{l} - \hat{f}(k_n)\, \Delta k \quad (139)$$

where $\Delta k = \dfrac{\pi}{l}$ is the distance between the neighbouring points representing the wave numbers in the spectrum [see formula (136) and Fig. 342]. Formula (115) representing $f(x)$ can therefore be rewritten as

$$f(x) = \sum_n c_n e^{ik_n x} = \sum_n \hat{f}(k_n)\, e^{ik_n x}\, \Delta k \quad (-l < x < l) \qquad (140)$$

Now suppose that l is very large. Then the spectrum becomes very "dense" and Δk very small. In the limit, as $l \to \infty$, sum (140) which is an integral sum turns into the corresponding integral, i.e. we obtain

$$f(x) = \int\limits_{-\infty}^{\infty} \hat{f}(k)\, e^{ikx}\, dk \quad (-\infty < x < \infty) \qquad (141)$$

In this representation the wave number k runs over all the values ranging from $-\infty$ to ∞, i.e. the discrete spectrum of wave numbers turns into a continuous spectrum in the limiting process as $l \to \infty$.

The first equality (139) shows that the amplitudes $c_n \to 0$ as $l \to \infty$. This means that in the limit each harmonic has a zero amplitude. But at the same time if we take an infinitesimal (but different from zero) interval of wave numbers between k and $k + dk$ we obtain the amplitude

$$dc = \hat{f}(k)\, dk \tag{142}$$

corresponding to the interval. Hence, the amplitude turns out to be distributed over the whole continuous spectrum of wave numbers, in the limit. This resembles the transition from a discrete model of a material body to its continuous model. Actually, in this transition we assume that the mass of each separate point is equal to zero and thus the total mass becomes continuously distributed over all the points with a certain density. By analogy, we can say that formula (142) describes the distribution of amplitudes of harmonics with density $\hat{f}(k)$. Thus, $f(k)$ is the density of the amplitude on an infinitesimal interval of wave numbers. The density is related to unit measure of length of the interval $[\hat{f}(k)$ is also referred to as the spectral density of the function $f(x)]$.

Formulas (138) and (141) express the so-called Fourier transformation*. Formula (138) defines the direct transformation and formula (141) the inverse transformation. We have deduced the formulas under the assumption that the function $f(x)$ is identically equal to zero outside a finite interval. Such functions are called finite**. A more extensive investigation shows that the formulas remain valid when integral (138) is understood as an improper integral. For the integral to be convergent. it is sufficient to impose the additional condition

$$\int_{-\infty}^{\infty} |f(x)|\, dx < \infty \tag{143}$$

Thus, to each function $f(x)$ satisfying condition (143) there corresponds its Fourier transform $\hat{f}(k)$ [which is the result, image. arising from the Fourier transformation applied to $f(x)]$. the transformation being defined by formula (138). Conversely. the function $f(x)$ is expressed in terms of its Fourier transform by formula (141) and is called the Fourier inverse transform [i.e. the inverse image, pre-

---

* The expression on the right-hand side of (141) is also called the Fourier integral.— *Tr.*

** The term a "finite function" should not be confused with the term a "bounded function" whose range is contained in some finite interval (although we sometimes say "finite" instead of "bounded"). To avoid the confusion we can use the term "a function of finite support" when speaking about functions identically vanishing outside an interval. the term taken from functional analysis.— *Tr.*

image, which is the result of the Fourier inverse transformation performed on the function $f(k)$].

If $f(x)$ is an even function we can use property 9, Sec. XIV.4, when computing integral (138), which yields

$$\hat{f}(k) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f(x) \cos kx \, dx - \frac{i}{2\pi} \int\limits_{-\infty}^{\infty} f(x) \sin kx \, dx =$$

$$= \frac{1}{\pi} \int\limits_{0}^{\infty} f(x) \cos kx \, dx$$

It follows that $\hat{f}(k)$ is also an even function in this case and hence, on the basis of formula (141), we obtain

$$f(x) = 2 \int\limits_{0}^{\infty} \hat{f}(k) \cos kx \, dk$$

These formulas define the so-called **Fourier cosine transform** and its inverse image. Similarly, if $f(x)$ is an odd function we arrive at the formulas defining the **Fourier sine transform** and its pre-image:

$$i\hat{f}(k) = \frac{1}{\pi} \int\limits_{0}^{\infty} f(x) \sin kx \, dx, \quad f(x) = 2 \int\limits_{0}^{\infty} i\hat{f}(k) \sin kx \, dk$$

By the way, in the last case the term "Fourier sine transform of a function $f(x)$" is usually applied to the function $i\hat{f}(k)$ instead of $\hat{f}(k)$.

If a function $f(x)$ is originally defined on the positive semi-axis $0 < x < \infty$ it can be extended onto the interval $-\infty < x < 0$ either as an even or as an odd function. Therefore considering both x and k to be positive we can use the cosine transform as well as the sine transform. But the images of these transformations will be different in the general case.

Let us take an example. Suppose $f(x)$ is an even function equal to 1 on the interval $-1 < x < 1$ and identically vanishing outside it. Then by the formula of the Fourier cosine transform we get

$$\hat{f}(k) = \frac{1}{\pi} \left(\int\limits_{0}^{1} 1 \cdot \cos kx \, dx + \int\limits_{1}^{\infty} 0 \cdot \cos kx \, dx \right) = \frac{\sin k}{\pi k}$$

Applying the inverse transformation we obtain

$$f(x) = 2 \int\limits_{0}^{\infty} \frac{\sin k}{\pi k} \cos kx \, dk \qquad (144)$$

As in the case of a Fourier series (see Sec. 25), if we substitute nume-rical values of x into the formula of the inverse Fourier transform we get the corresponding values of $f(x)$ at all the points of continuity of f and the values $\frac{1}{2}[f(x-0)+f(x+0)]$ at all points where f has a finite jump. In particular, substituting the value $x = 0$ into (144) for which the function f is continuous we get

$$1 = 2\int_0^\infty \frac{\sin k}{\pi k}\,dk, \quad \text{which implies} \quad \int_0^\infty \frac{\sin k}{k}\,dk = \frac{\pi}{2}$$

An integral formula equivalent to the formulas of the Fourier transforms was obtained by Fourier in 1811.

33. Properties of Fourier Transforms. A Fourier transform possesses many useful properties. We are going to enumerate some of them here. First of all, it is clear that a Fourier transformation can be interpreted as an operator (see Sec. XIV.26) for which the function $f(x)$ is the inverse image (pre-image) and the function $\hat{f}(k)$ is the image.

1. The Fourier operator is linear, that is

$$\widehat{(f_1+f_2)} = \hat{f}_1 + \hat{f}_2, \quad \widehat{\alpha f} = \alpha \hat{f} \quad (\alpha = \text{const}) \tag{145}$$

This is directly implied by formula (138) and by the fact that the integration is a linear operation.

Formula (145) implies, in particular, that if f depends not only on x but also on a parameter t the function $\hat{f}$ is dependent on the parameter as well, and we have

$$\frac{\widehat{f_{t+\Delta t} - f_t}}{\Delta t} = \frac{\hat{f}_{t+\Delta t} - \hat{f}_t}{\Delta t}$$

Passing to the limit, as $\Delta t \to 0$, we obtain

$$\frac{\widehat{\partial f}}{\partial t} = \frac{\partial \hat{f}}{\partial t}$$

Consequently, the derivative with respect to a parameter of the Fourier inverse transform is transformed into the derivative with respect to the parameter of the Fourier transform. By the way, the method of proving the property indicates that this property is common to all the linear operators.

2. The differentiation of the function f with respect to x results in the multiplication of its Fourier transform $\hat{f}$ by ik. Indeed, the

Fourier inverse transform of the function $f'(x)$ is

$$\frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f'(x)\, e^{-ikx}\, dx = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{-ikx}\, df(x) =$$

$$= \frac{1}{2\pi}\, e^{-ikx} f(x)\, \Big|_{x=-\infty}^{x=\infty} - \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f(x)\,(-ik)\, e^{-ikx}\, dx$$

(we have performed integration by parts here). But condition (143) indicates that $f(\pm\infty) = 0$ and therefore the first summand on the right-hand side vanishes. The second summand is equal to

$$\frac{ik}{2\pi} \int\limits_{-\infty}^{\infty} f(x)\, e^{-ikx}\, dx = ik\hat{f}(k)$$

which is what we set out to prove.

Transforming formula (141) in a similar way we can prove that if the function $\hat{f}$ is differentiated its Fourier inverse transform f is multiplied by $-ik$.

3. If the function $\hat{f}(k)$ is the Fourier transform of the function $f(x)$ the function $\frac{1}{\alpha}\, \hat{f}\left(\frac{k}{\alpha}\right)$ $(\alpha = \text{const} > 0)$ is the Fourier transform of $f(\alpha x)$. Actually, performing the change of variable $\alpha x = s$ we obtain

$$\frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f(\alpha x)\, e^{-ikx}\, dx = \frac{1}{\alpha}\, \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f(s)\, e^{-i\frac{k}{\alpha}s}\, ds = \frac{1}{\alpha}\, \hat{f}\left(\frac{k}{\alpha}\right)$$

Therefore, if the graph of the Fourier inverse transform is stretched α-fold along the x-axis, the graph of the Fourier transform is contracted α-fold along the k-axis and vice versa. This means that we cannot simultaneously localize (that is concentrate at a certain point of the corresponding axis) both a function which serves as a Fourier inverse transform and its spectral density. This is the so-called *uncertainty principle* which has many applications in physics.

4. If the function $f(x)$ is shifted by $\beta = \text{const}$ along the x-axis (i.e. its graph is shifted by the distance β), its Fourier transform is multiplied by $e^{-i\beta k}$. In fact, making the substitution $x - \beta = s$ we obtain

$$\frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f(x-\beta)\, e^{-ikx}\, dx = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f(s)\, e^{-iks} e^{-ik\beta}\, ds = e^{-ik\beta}\hat{f}(k)$$

Conversely, if the transform is shifted by β along the k-axis the inverse transform is multiplied by $e^{i\beta x}$.

5. *Parseval's Theorem*. If we apply formula (123) to series (115) we obtain

$$\int_{-l}^{l} |f(x)|^2\, dx = 2l \sum_{n=-\infty}^{\infty} |c_n|^2$$

Taking advantage of formula (139) we deduce from the last relation the equality

$$\int_{-l}^{l} |f(x)|^2\, dx = 2l \sum_{n=-\infty}^{\infty} |\hat{f}(k_n)|^2 (\Delta k)^2 = 2\pi \sum_{n} |\hat{f}(k_n)|^2 \Delta k$$

Passing to the limit as $l \to \infty$ we obtain, by analogy with Sec. 32, the relation

$$\int_{-\infty}^{\infty} |f(x)|^2\, dx = 2\pi \int_{-\infty}^{\infty} |\hat{f}(k)|^2\, dk$$

It is this relation that is called **Parseval's theorem for the Fourier transform.**

34. Application to Oscillations of Infinite String. The Fourier integral transformation is applied to solving some problems of mathematical physics for infinite media. We shall illustrate the solution of equation (131) in the case of an infinite string, i.e. when $-\infty < x < \infty$. Hence, there are no boundary conditions in this problem and it is only the initial conditions that define the sought-for solution. For the sake of simplicity, let us put $\psi(x) \equiv 0$ in the initial conditions (133). We denote by $\hat{u}(k,\, t)$ the Fourier transform of the solution $u(x,\, t)$ for any fixed value of $t \geqslant 0$. Passing to the Fourier transforms of the left-hand and right-hand sides of equation (131) and taking advantage of properties 1 and 2 in Sec. 33 we obtain

$$\frac{\partial^2 \hat{u}}{\partial t^2} = a^2 (ik)^2\, \hat{u} = -a^2 k^2 \hat{u}$$

For any fixed k, this is an ordinary linear differential equation with constant coefficients which can be solved by means of the standard methods given in Sec. XV.17. Thus we obtain

$$\hat{u} = C_1(k)\, e^{-iakt} + C_2(k)\, e^{iakt} \tag{146}$$

Now we take the Fourier transforms of the initial conditions (133):

$$\hat{u}\,|_{t=0} = \hat{\varphi}(k) \quad \text{and} \quad \frac{\partial \hat{u}}{\partial t}\bigg|_{t=0} = 0$$

Consequently, formula (146) results in

$$\hat{u} = \frac{1}{2}\, \hat{\varphi}(k)\, e^{-iatk} + \frac{1}{2}\, \hat{\varphi}(k)\, e^{iatk}$$

(verify the calculations!). On the basis of properties 1 and 4 in Sec. 33, we now return to the inverse image:

$$u = \frac{1}{2}\,\varphi\,(x - at) + \frac{1}{2}\,\varphi\,(x + at)$$

It is the last formula that yields the sought-for solution. The meaning of the formula is quite simple: the initial deflection from the equilibrium state is divided into two equal parts; one of the parts is shifted at the moment t by the distance at in the positive direction of the x-axis whereas the other is shifted by the same distance in the opposite direction. In other words, the two waves propagate along the string with the speed a in the positive and negative directions without changing their initial form. At each moment of time we watch the result of the superposition of the waves. Thus, we see what is the physical meaning of the constant a entering into equation (131): it is equal to the speed at which an initial perturbation propagates along the string.

Elements of the Theory of Probability

§ 1. Random Events and Their Probabilities

1. Random Events. The theory of probability deals with random events. The notion of an event is a basic one, and it is rather difficult to give its comprehensive definition.

For the aims of our course it will be sufficient to regard as an event everything that may or may not occur when a certain set of conditions is realized. Every realization of this kind is called a **trial**. For instance, when tossing a coin we can consider the fact that it shows heads to be an event. In this case tossing the coin serves as a trial. We can regard it as an event when an article randomly selected from a lot containing a number of manufactured articles turns out to be defective. In this example sampling a unit from the lot is a trial. But a trial, as it is understood in the theory of probability, must not necessarily be connected with human activities. For example, if we consider it to be an event that it will rain in a certain place on a certain day, the fact that this day comes should be regarded as a trial.

A characteristic feature of a **random event** is that it may not necessarily occur when a trial is realized. This distinguishes a random event from a deterministic one which inevitably occurs. The randomness of an event is connected with the fact that many concomitant factors which are essential for the outcome of a trial may not be given. The incompleteness of information can sometimes be intrinsic (for instance, in games of chance or in warfare) or can result from the inaccessibility of some kind of information at the present level of the development of science (for example, in problems of weather forecast). The assumption that the outcomes of individual trials cannot be predicted is taken as a basic principle in quantum mechanics, genetics and some other sciences. Besides, there are some cases in which the exact prediction of the outcomes of certain trials is possible but not advantageous when it requires unnecessary expenditure connected with additional precision measurements and the like.

Regularities of random events appear in *mass-scale phenomena* when trials are repeated a large number of times. For instance, we cannot predict the result of a single toss of a coin because it can come up heads or tails. Nobody will find it very strange if we have heads twice when we toss a coin ten times. But if we get only 200 heads after the coin has been tossed 1000 times we have every reason to say that there is something wrong with the coin or with tossing. Indeed, if the conditions are equal neither heads nor tails have any advantage and therefore they must appear approximately equal number of times. Of course, when tossing an "honest" coin 1000 times we may not necessarily have heads exactly 500 times; we can have them 490 or 525 times or so but not 200 times! Similarly, if we examine a single unit selected at random from a lot we cannot have a good judgement on the quality of the lot. This can be done only after a sufficiently large number of repeated trials have been performed or, as we say, when we have sufficiently large sample size. Thus, specifying what was said at the beginning of this section, we can say that the theory of probability deals with random events which occur in mass-scale phenomena when the corresponding set of conditions is realized a large number of times.

There are two possible ways of understanding the repetition of trials. For instance, we can toss one and the same coin 1000 times but we can also independently toss 1000 similar coins at different instances of time or even simultaneously. Both possibilities are quite equivalent. and further we shall not distinguish between them.

2. Probability. In everyday life we often say that a certain event is highly probable whereas some other event is improbable. Of course, in case the corresponding trials can be repeated many times these assertions mean that the former event will occur frequently and the latter will occur seldom.

An important feature of the theory of probability is that it not only indicates that the probability of an event is high or low but also attributes an exact numerical value to it. Hence, the probability of an event is considered to be a numerical value which characterizes the frequency of the occurrences of the event in a large number of repeated trials.

Suppose that a coin was tossed 1000 times and that it came up heads 490 times. Then the ratio $\frac{490}{1000} = 0.49$ is said to be the **relative frequency** of the coin coming up heads in the given series of trials. Let the coin be tossed 10,000 times and let heads appear 5027 times; then the relative frequency is equal to 0.5027. It is clear that if the coin is symmetric and if the number of trials increases the relative frequency of heads must approach 0.5 because neither of the faces of the coin has any advantage over the other. It is the number

0.5 that is called the **probability** of heads when the coin is being tossed.

In the general case the definition is stated similarly. Denote a random event by A. Suppose that the event occurred N_A times in a series of N independent trials. Then the ratio $\frac{N_A}{N}$ is called the relative frequency of the event A in the given series of trials.

The limit

$$P\{A\} = \lim_{N\to\infty} \frac{N_A}{N}$$

to which the relative frequency of the event A tends when the number of trials is increased unlimitedly is called the probability of the random event A.

Hence, if the number of trials is sufficiently large the relative frequency of an event can be approximately taken as its probability. This fact implies a method of empirical calculation of probabilities when it is difficult to find them theoretically. Let us consider an example. The probability that a new-born child will be a boy is known with a great accuracy from the statistics of human population. The probability equals 0.512 although from time to time there appears a deviation from this value. Knowing this probability we cannot predict whether a new-born child will be a boy or a girl for each concrete case. We can only say that the probability of the child being a boy is a little higher, i.e. the birth of a boy is a little more probable. But nevertheless we can assert that the number of boys among a million of new-born children will be close to 512,000 (in Sec. 20 we shall work out some methods for estimating the degree of this closeness).

There are some cases when the probability can be calculated by means of figuring the number of **favourable trial outcomes (favourable cases)**. We shall illustrate this method by taking a concrete example. Suppose that we toss a die on whose faces sums of points from 1 to 6 are marked. Let it be necessary to find the probability of a throw giving a sum of points divisible by 3. Imagine that we have made a large number N of throws. Let N_k denote the number of occurrences of k points ($k = 1, 2, 3, 4, 5$ and 6). Then we have $N_1 + N_2 + \ldots$ $\ldots + N_6 = N$. that is $\frac{N_1}{N} + \frac{N_2}{N} + \ldots + \frac{N_6}{N} = 1$. But neither of the faces having any advantage over the others, all the six fractions are approximately equal to $\frac{1}{6}$ if N is sufficiently large, and in the limit they become exactly equal to each other when $N \to \infty$. Hence, in the limit the fractions are equal to $\frac{1}{6}$. But a sum of points is multiple of 3 if it equals three or six, and hence the number of

favourable outcomes is equal to $N_3 + N_6$. The relative frequency of the event is equal to

$$\frac{N_3 + N_6}{N} = \frac{N_3}{N} + \frac{N_6}{N} \xrightarrow[N \to \infty]{} 2 \cdot \frac{1}{6} = \frac{1}{3}$$

This is just the sought-for probability. The result thus obtained can be formulated briefly as follows: we have six possible outcomes in throwing a die which correspond to the range of possible sums of points. There are two outcomes among them which are favourable to the event in question, namely throwing three and six. The other cases are unfavourable. Hence, the probability of the event equals $\frac{2}{6} = \frac{1}{3}$.

The general scheme of such calculations can be described in the following way. Suppose that a trial can result in exactly one of n possible outcomes, these outcomes being equally probable (in such circumstances we also say that we have n equally probable possible cases). Let us consider an event A which occurs when q of these outcomes appear and does not occur when the other $n - q$ outcomes appear. We call these q outcomes **favourable to the event** A whereas the other $n - q$ cases are called **unfavourable to** A. Then, reasoning as in the preceding paragraph, we conclude that

$$P\{A\} = \frac{q}{n}$$

Thus, the probability of an event is equal to the ratio of the number of trial outcomes favourable to the event to the number of all possible outcomes.

Using the scheme of favourable cases we can easily find the probability of winning a prize for an owner of a lottery ticket (for this purpose the total number of prizes should be divided by the number of tickets). Many other similar probabilities can be found in a similar way. In performing such calculations we must be sure that the outcomes of trials are equally possible, i.e. equally probable. For instance, it would be wrong to reason in the following way: the sum of points obtained when tossing the die can be either divisible by three or indivisible and therefore there is one favourable case among the two for getting a sum multiple of three, and hence the sought-for probability equals $\frac{1}{2}$ (where does the mistake lie in this argument?). Of course, we always idealize reality when we consider some outcomes to be equally possible, and therefore this assumption only approximately holds in concrete problems, with a certain accuracy. Such an approach is justified only when there is symmetry in trials under consideration.

There are various modifications of the scheme of figuring the number of favourable cases. Let us consider an example illustrating

one of the variants of the method. Suppose we have a homogeneous spherical ball with surface area S. Let a portion of the surface having an area S_0 be blackened. Let the ball be thrown at random on a horizontal plane and let it be necessary to calculate the probability that the ball strikes the plane with the portion S_0 of its surface. To perform the calculation we imagine that the whole surface is divided into small parts of equal areas dS. Then the point of impact can belong to any of these parts with an equal probability. But the total number of these parts is equal to $\frac{S}{dS}$, and the number of the parts belonging to the portion S_0 is equal to $\frac{S_0}{dS}$. Hence, there are $\frac{S_0}{dS}$ cases favourable to the event in question among the total number of cases $\frac{S}{dS}$, and thus the sought-for probability is equal to $\frac{S_0}{dS} : \frac{S}{dS} =$ $= \frac{S_0}{S}$. If we pose the same problem for an ellipsoid instead of the ball the solution will depend not only on the area but also on the disposition of the blackened portion on the surface of the ellipsoid. The solution involves integration, and we leave it to the reader.

In conclusion we note that in everyday life the term "probability" is sometimes applied to such events which cannot be repeated, even mentally. For instance, we sometimes speak about the probability of whether there exists life on Mars and the like. In such cases it would be better to speak about estimating the likelihood of a hypothesis. The likelihood theory is not thoroughly developed at present.

3. Basic Properties of Probabilities.

1. The probability of any event A is a dimensionless quantity whose numerical value lies between the limits 0 and 1:

$$0 \leqslant P\{A\} \leqslant 1$$

The property immediately follows from the definition of probability given in Sec. 2. The definition also indicates that the greater $P\{A\}$, the greater the possibility of the occurrence of the event, i.e. the greater its probability understood in everyday sense.

2. The probability of a **certain (sure) event**, that is of an event which unavoidably occurs, is equal to unity. Thus, we regard a certain, deterministic event as a special case of a random event (this resembles our arguments in Sec. I.5 where we considered a constant quantity to be a special case of a variable quantity). Further, the probability of an **impossible event** is equal to zero.

In the case of a finite number of possible outcomes of trials the converse assertions are also true. Namely, if the probability is equal to unity (zero) the event is certain (impossible). But in the general case these assertions are no longer true. For instance, our discussion

in Sec. 2 shows that the probability that a ball thrown at random will strike a plane with a point which is set beforehand is equal to zero. At the same time such an event is not impossible (theoretically) because its occurrence does not contradict the laws of mechanics. But of course the event is *practically impossible.*

3. The sum of probabilities of any event and its **opposite event** is always equal to unity. We say that two events are **contrary** or **opposite** to each other if the occurrence of one of them is equivalent to the non-occurrence of the other. In other words, each of the two contrary events is the negation of the other. If the probability of hitting a target under certain conditions is equal to 0.2 the probability of failing to hit the target under the same conditions is equal to 0.8. To prove this assertion in the general case we denote two contrary events by A and $\overline{A}$. Let N trials be made and let the event A occur N_A times and the event $\overline{A}$ occur $N_{\overline{A}}$ times. Then it is evident that $N_A + N_{\overline{A}} = N$ which implies $\frac{N_A}{N} + \frac{N_{\overline{A}}}{N} = 1$. Passing to the limit for $N \rightarrow \infty$ we find that

$$\mathbf{P}\{A\} + \mathbf{P}\{\overline{A}\} = 1 \tag{1}$$

4. We can similarly prove a more general assertion: if a trial results in the necessary occurrence of one and only one event belonging to a group of events $A_1, A_2, \ldots, A_k$ we have

$$\mathbf{P}\{A_1\} + \mathbf{P}\{A_2\} + \ldots + \mathbf{P}\{A_k\} = 1 \tag{2}$$

5. We now consider two events A and B such that each of them may or may not occur when one and the same trial is performed. Suppose that N such trials have been made. Let $N_{A\,and\,B}$ designate the number of trials in which both events occurred and let $N_{A\,and\,\overline{B}}$ be the number of trials in which A occurred and B did not occur and so on. Using this notation we can write

$$N = N_{A\,and\,B} + N_{A\,and\,\overline{B}} + N_{\overline{A}\,and\,B} + N_{\overline{A}\,and\,\overline{B}}$$

Besides. for the total number of trials in which the event A occurred and for the total number of trials in which the event B occurred we can write

$$N_A = N_{A\,and\,B} + N_{A\,and\,\overline{B}} \quad \text{and} \quad N_B = N_{A\,and\,B} \quad + \atop + N_{\overline{A}\,and\,B} \tag{3}$$

Furthermore, let us denote the number of trials in which at least one of the events A and B occurred by $N_{A\,or\,B}$. Then we have

$$N_{A\,or\,B} = N_{A\,and\,B} + N_{A\,and\,\overline{B}} + N_{\overline{A}\,and\,B} \tag{4}$$

Formulas (3) and (4) imply

$$\frac{N_{A \ or \ B}}{N} = \frac{N_A}{N} + \frac{N_B}{N} - \frac{N_{A \ and \ B}}{N}$$

Passing to the limit, as $N \to \infty$, we arrive at the formula

$$P \{A \ or \ B\} = P \{A\} + P \{B\} - P \{A \ and \ B\}$$

in which the sense of the notation is quite clear.

In particular, if the events A and B are **mutually exclusive**, that is such that they cannot occur simultaneously, we obtain the following theorem of addition of probabilities (addition rule of probability theory):

$$P \{A \ or \ B\} = P \{A\} + P \{B\} \quad \text{(where } A \text{ and } B$$
$$\text{are mutually exclusive)}$$

The following more general rule is proved in a similar way: if the events $A_1, A_2, \ldots, A_k$ are **pairwise mutually exclusive** we have

$$P \{A_1 \ or \ A_2 \ldots or \ A_k\} = P \{A_1\} + P \{A_2\} + \ldots + P \{A_k\}$$
$$\tag{5}$$

4. **Theorem of Multiplication of Probabilities.** Let A and B be two events. Then the **conditional probability** $P \{A \mid B\}$ of the event A relative to the hypothesis that the event B has occurred is the probability of the event A calculated on the condition that the event B has taken place. Therefore, when calculating this probability by means of the corresponding relative frequency (see Sec. 2), we must take into account only those trials whose outcomes resulted in the occurrence of the event B:

$$P \{A \mid B\} = \lim_{N \to \infty} \frac{N_{A \ and \ B}}{N_B}$$

For instance, suppose we are given two urns. Let the first urn contain three black balls and one white ball and the second one contain one black ball and three white balls. Suppose that we randomly select one of the urns and draw a ball from it at random. What is the probability that the ball will be black? The obvious symmetry of the possible outcomes indicates that $P \{A_{black}\} = \frac{1}{2}$ where A_{black} is the event consisting in the occurrence of a black ball. We now suppose that it is known that we have selected the first urn. Let us denote this event (that is selecting the first urn) as B_1. Then it is apparent that the conditional probability $P \{A_{black} \mid B_1\} = \frac{3}{4}$.

Taking the simple formula

$$\frac{N_{A \, and \, B}}{N} = \frac{N_B}{N} \, \frac{N_{A \, and \, B}}{N_B}$$

and passing to the limit, as $N \to \infty$, we obtain the following multiplication rule of probability theory:

$$P\{A \, and \, B\} = P\{B\} \, P\{A \mid B\} = P\{A\} \, P\{B \mid A\} \qquad (6)$$

(the last expression has been obtained by interchanging the roles of A and B).

Thus, the probability that two events take place simultaneously is equal to the product of the probability of one of them by the conditional probability of the other provided that the first event has occurred.

Formula (6) becomes especially simple when the events A and B are independent. We call two events independent if any information concerning the occurrence or non-occurrence of one of them does not affect the probability of the other. Thus, in this case we have

$$P\{A \mid B\} = P\{A\}, \quad P\{A \mid \overline{B}\} = P\{A\},$$
$$P\{B \mid A\} = P\{B\}, \quad P\{B \mid \overline{A}\} = P\{B\}$$

(By the way, on the basis of equalities (1), (5) and (6), we can easily conclude that each of the above relations implies the other three.) Formula (6), for independent events, turns into

$$P\{A \, and \, B\} = P\{A\} \, P\{B\} \qquad (7)$$
(where A and B are independent)

Formula (7) can be readily extended to the case of an arbitrary number of independent events, that is events such that the information concerning the occurrence or non-occurrence of any group of these events does not affect the probabilities of the others. For example, if A, B and C are independent events we have

$$P\{A \, and \, B \, and \, C\} = P\{A \, and \, (B \, and \, C)\} =$$
$$= P\{A\} \, P\{B \, and \, C\} = P\{A\} \, P\{B\} \, P\{C\} \qquad (8)$$

The above rules enable us to calculate probabilities for some simple problems. Let us take an example. Suppose there are three shots and each of them fires at a target once. Let the first of them hit the target with a probability of 0.2, the second with a probability of 0.3 and the third with 0.5. What is the probability that the target will be hit at least once? If we denote the probability of kth shot hitting the target as A_k ($k = 1$, 2, 3) we can say that we are interested in the probability $P\{A_1 \, or \, A_2 \, or \, A_3\}$. We cannot apply formula (5) here because the events A_k are not mutually

exclusive since the target can be simultaneously hit by two or three shots. Therefore in this case it is easier to calculate the probability that all the shots miss the target because these opposite events are independent. Thus, by formulas (1) and (8), we obtain

$$P\ \{A_1\ or\ A_2\ or\ A_3\} = 1 - P\ \{\overline{A}_1\ and\ \overline{A}_2\ and\ \overline{A}_3\} =$$
$$= 1 - P\ \{\overline{A}_1\}\ P\ \{\overline{A}_2\}\ P\ \{\overline{A}_3\} = 1 - 0.8 \times 0.7 \times 0.5 = 0.72$$

Let us consider one more example. Suppose that we randomly draw two balls in succession from an urn containing three black balls and one white ball. What is the probability that both balls will be black? There can be two variants of the problem. Namely, we can consider *sampling with replacement*. In our case this means that we consider drawing a ball with replacement which means that the first ball drawn from the urn is replaced in the urn after its colour has been noted and before the next drawing is made. Hence, the case when the same ball will be drawn a second time is not excluded here. Formula (7) is obviously applicable here, and thus we see that the sought-for probability is equal to $\frac{3}{4} \cdot \frac{3}{4} =$ $= \frac{9}{16}$. But if we consider *drawing without replacement*, that is if the first ball drawn from the urn is not replaced in the urn after its colour has been examined, then this ball does not take part in the second sampling, and therefore the sought-for probability is calculated by formula (6) which yields $\frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2}$.

5. Theorem of Total Probability. We shall begin with an example. Let there be three urns. The first urn contains three black balls and one white ball, the second contains one black ball and three white balls and the third only three black balls. Suppose we randomly selected one of the urns (with equal probability) and then drew a ball from the urn at random. What is the probability of the ball being black? If we drew a black ball this obviously indicates that we either selected the first urn and drew a black ball from it or did the same with the second or with the third urn. All these three variants are pairwise mutually exclusive. By formula (7), the probability of the first variant taking place is equal to $\frac{1}{3} \cdot \frac{3}{4}$, the probability of the second variant is equal to $\frac{1}{3} \cdot \frac{1}{4}$ and the probability of the third one is equal to $\frac{1}{3} \cdot 1$. Hence, the probability of the occurrence of one of the variants is equal, by formula (5), to

$$\frac{1}{3} \cdot \frac{3}{4} + \frac{1}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot 1 = \frac{2}{3}$$

This is just the sought-for probability.

Now let us turn to the general case. Suppose that the result of a trial is the occurrence of one and only one of the k events B_1, B_2, ..., B_k which are pairwise mutually exclusive (in the previous example the role of such events was played by the selections of one of the urns). Besides, let us consider an event A (in the above example the role of A was played by the drawing of a black ball). We can regard A as being equivalent to the event consisting in the occurrence of B_1 and A, or of B_2 and A, · or of B_3 and A and so on. All the last variants being mutually exclusive, formula (5) implies

$$\mathbf{P}\{A\} = \mathbf{P}\{(B_1 \text{ and } A) \quad or \quad (B_2 \text{ and } A) \dots or \ (B_k \text{ and } A)\} =$$
$$= \mathbf{P}\{B_1 \text{ and } A\} + \mathbf{P}\{B_2 \text{ and } A\} + \dots + \mathbf{P}\{B_k \text{ and } A\}$$

From this, by formula (6), we finally deduce

$$\mathbf{P}\{A\} = \mathbf{P}\{A \mid B_1\}\ \mathbf{P}\{B_1\} + \mathbf{P}\{A \mid B_2\}\ \ \mathbf{P}\{B_2\} + \dots +$$
$$+ \mathbf{P}\{A \mid B_k\}\ \mathbf{P}\{B_k\} \tag{9}$$

This formula is called the **formula of total probability (partition formula)**. It can be applied to problems similar to the one considered in the foregoing paragraph.

6. **Formulas for the Probability of Hypotheses.** We begin with the above example of the three urns again. Suppose that we know the distribution of the balls in the urns and that the urns themselves are indistinguishable. This means that when we select one of the urns at random we do not know which of the urns has been selected. Then, considering the three hypotheses that we have selected the first urn or the second urn or the third one we conclude that they are all equally probable, that is the probability of each of the hypotheses is equal to $\frac{1}{3}$. Now let us draw a ball at random from the urn we have selected and let the ball turn out to be white. Then we should reappraise the probabilities of the hypotheses. For instance, after the drawing of a white ball, it becomes clear that the urn we have selected cannot be the third one and that it is more probable that we have selected the second urn than the first (why?). The probabilities calculated before the performance of the experiment (i.e. before drawing a ball) are called **a priori probabilities**, and the reappraised probabilities are called **a posteriori probabilities** (the term *a priori* originates from Latin and means presumptive, and *a posteriori* is the reverse of *a priori*). Now, how can we find these reappraised probabilities?

Let us take the general case. Let there be several hypotheses H_1, H_2, ..., H_k and let it be known that one and only one of them holds. Let the a priori probabilities of the hypotheses be equal to $\mathbf{P}\{H_1\}$. $\mathbf{P}\{H_2\}$, ..., $\mathbf{P}\{H_k\}$, respectively. Suppose that the conditional probabilities $\mathbf{P}\{A \mid H_i\}$ $(i = 1, 2, ..., k)$ of an event

A relative to each of the hypotheses are known. Then the a posteriori probabilities of the hypotheses are nothing but the probabilities $P\{H_i \mid A\}$ ($i = 1, 2, \ldots k$). To calculate them we write, on the basis of (6), the relations

$$P\{A\}\, P\{H_i \mid A\} = P\{H_i\}\, P\{A \mid H_i\}$$

and then, applying formula (9), deduce

$$P\{H_i \mid A\} = \frac{P\{A \mid H_i\}\, P\{H_i\}}{P\{A \mid H_1\}\, P\{H_1\} + P\{A \mid H_2\}\, P\{H_2\} + \ldots + P\{A \mid H_h\}\, P\{H_h\}}$$

$$i = 1, \ldots, k$$

These are the sought-for formulas for the probability of hypotheses (Bayes' theorem). Let the reader apply the formulas to verify that the reappraised probabilities in the problem considered in the preceding paragraph are equal to $\frac{1}{4}$, $\frac{3}{4}$ and 0, respectively.

7. Disregarding Low-Probability Events. We see that the methods of the theory of probability enable us to calculate the probabilities of various events. How can we utilize these results? One can hardly be satisfied with the answer that a given event will either occur or not.

There is an approach to the problem which is typical of applied mathematics. It is based on the idea that if the probability of an event A under consideration is sufficiently small, that is if $P\{A\} <$ $< \varepsilon$ where ε is a sufficiently small positive number, we can approximately put $P\{A\} = 0$ and thus consider the event A to be practically impossible. In such a case we simply disregard the possibility that A may occur. Of course, this does not exclude the theoretical possibility of the occurrence of A, and therefore the prediction that A will not occur may turn out to be wrong. But the smaller ε, the rarer the occurrences of the event.

But how can we choose ε? There are various traditions concerning this question in different divisions of applied mathematics. If there is nothing dangerous in the occurrence of the event A, that is if the error introduced by the incorrect prediction can be easily corrected, we can put $\varepsilon = 0.1$. This means that in the long run approximately 10 per cent of predictions will be false. But if a higher reliability is not connected with essential difficulties we usually put $\varepsilon = 0.01$. For instance, if we toss a coin 100 times the meaning of the choice of $\varepsilon = 0.01$ is that we disregard the possibility of such events as the coin coming up heads seven times in succession because $2^7 \approx 100$. For still more accurate predictions we can put $\varepsilon = 0.001$; then the average frequency of incorrect predictions will be about one per thousand and so on. The smaller ε, the more accurate the prediction. But at the same time it is more difficult to guarantee such an accuracy when ε is decreased. The accuracy

should be particularly great if an incorrect prediction may be connected with casualties. In such cases we sometimes cannot rely upon probabilistic inferences, and then we have to resort to deterministic ones.

In Sec. 14 we shall discuss some methods of choosing criteria (i.e. choosing ε) according to which events can be considered to be practically impossible.

§ 2. Random Variables

8. **Definitions.** A **random variable** is a variable quantity which randomly assumes a certain numerical value resulting from the outcome of a trial. This value depends on chance and, generally speaking, varies as the trials are repeated.

Examples of random variables are the number of students attending a lecture, the length of a manufactured article taken from a lot, the duration of life of a person and so on.

Like every quantity (see Sec. I.5), a random variable can be **discrete** or **continuous**. For instance, the first random variable in the above examples is discrete whereas the other two are continuous. It is essential here that even before a trial has been made we know that the possible values of the number of students are integral, whereas it is impossible to set beforehand the possible discrete values of the lengths of the articles.

To obtain a representation of a discrete random variable we can enumerate all its possible values and indicate the probabilities with which these values are assumed. This results in a table of the following form:

<div align="center">DISCRETE RANDOM VARIABLE ξ</div>

values of ξ	x_1	x_2	x_3	. . . .
probabilities	P_1	P_2	P_3	. . . .

(10)

Such a table can be finite or infinite (theoretically). It is apparent that all the probabilities P_k must be non-negative and their sum, according to formula (2), must be equal to unity. In the special case when there is only one possible value it must be assumed with probability 1, that is the variable in question necessarily assumes this value. Thus, in such a case we have a deterministic quantity.

A continuous random variable ξ can assume all the numerical values or all the values belonging to some interval (or to a system of intervals). But the probability that such a random variable will exactly take on any value x set beforehand is equal to zero. (This

situation is similar to that of a continuously distributed mass when the mass of any separate point is considered to be equal to zero.) But we can speak about the probability that the random variable ξ will assume a value belonging to a given interval of the x-axis. The probability of the value of ξ to fall in an infinitesimal interval from x to $x + dx$ is also infinitesimal; it is directly proportional to dx and depends on x. Hence, this probability is equal to an expression of the form $p(x)\,dx$ where $p(x)$ is the so-called **probability density function** (the **density of probability distribution** or the **frequency function** of the probability distribution). This function completely characterizes the random variable ξ. Apparently. we always have $p(x) \geqslant 0$ ($-\infty < x < \infty$). Formula (5) shows that the probability that the value assumed by the random variable ξ belongs to an interval $a \leqslant x \leqslant b$ is equal to $\int_a^b p(x)\,dx$. Hence, by formula (2), there must be

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \tag{11}$$

The expression $p(x)\,dx$ is called the *element of the probability distribution*. Henceforward we shall write integrals taken from $-\infty$ to ∞ without indicating the limits of integration; for instance, formula (11) will be put down as $\int p(x)\,dx = 1$. This will not lead to any misunderstandings because we shall not deal with indefinite integrals in this chapter. (By the way, the sign $\int$ is rarely used in mathematical applications for denoting indefinite integrals. It usually designates definite integrals for which the limits of integration are implied by the corresponding physical or mathematical meaning of the integrals. For instance, we sometimes mean that the sign $\int$ designates definite integrals taken over maximal ranges of variation of the corresponding variables of integration.)

Using the notion of the delta function (see Sec. XIV.25) we can also introduce the probability density function for a discrete random variable. For instance, if such a random variable is represented by a table of form (10) we have

$$p(x) = P_1\delta(x - x_1) + P_2\delta(x - x_2) + P_3\delta(x - x_3) + \ldots$$

Delta functions also enable us to consider the density of probability distribution for a random variable of a mixed (continuous-discrete) type. In the general case the representation of a random variable is equivalent to the construction of a non-negative measure in the

straight line (see Sec. XVI.19) such that the total measure of the whole straight line is equal to unity. Then, such a probability measure being given, the probability that the value of the random variable will fall in an interval is equal to the measure of the interval.

9. Examples of Discrete Random Variables. We now consider a random event A with $\mathbf{P}\{A\} = P$. Suppose that we have made one trial. Then how many times can the event A occur? Evidently, this number is equal either to 1 or to 0. Hence, we have obtained a random variable which can assume only two values, namely the value 1 with probability P and the value 0 with probability $1 - P$ [see formula (1)].

Now let the trials be performed several times. For definiteness, let there be three trials. Suppose that the event A has occurred ν times in these trials. Then ν can be regarded as a random variable whose possible values are 0, 1, 2 and 3. Let us compute the probabilities of these values. If $\nu = 0$ the event A does not occur in all the three trials. The trials being independent, the probability that $\nu = 0$ can be found by formula (8). This results in the probability equal to $(1 - P)^3$. The value $\nu = 1$ can be obtained in the following three variants: the event A occurs in the first (second or third) trial and does not occur in the other two trials. The probability of each variant is again found by formula (8) which yields the result $P(1 - P)^2$. Therefore, according to formula (5), the probability that we shall have one of the variants is equal to $3P(1 - P)^2$. The cases $\nu = 2$ and $\nu = 3$ are treated similarly, and thus we arrive at the following table:

values of ν	0	1	2	3
probabilities	$(1-P)^3$	$3P(1-P)^2$	$3P^2(1-P)$	P^3

(Let the reader check up that the sum of the probabilities thus obtained is equal to unity!)

The general case of n trials is investigated in like manner. Let again ν be the number of the occurrences of the event A. Then ν is a random variable for which we obtain the table

values of ν	0	1	2	...	n
probabilities	$(1-P)^n$	$\binom{n}{1} P(1-P)^{n-1}$	$\binom{n}{2} P^2(1-P)^{n-2}$	...	P^n

Here $\binom{n}{k} = \frac{n(n-1)\ldots(n-k+1)}{k!}$ is a binomial coefficient which is equal to the number of possible cases in which the event A occurs exactly k times, i.e. it is equal to the number of combinations of k elements from n. The set of probabilities collected in the above table is called the **law of binomial probability distribution** (or, briefly, the **binomial distribution**).

Example. A coin is tossed six times. What is the probability that it will come up heads exactly three times?

Answer: $\binom{6}{3} \frac{1}{2^3} \left(1 - \frac{1}{2}\right)^3 = \frac{5}{16}$.

Now let us investigate the behaviour of the binomial distribution when n, the number of trials, is very large whereas the probability of the event A is very small so that there is a relation $Pn = \alpha$ where α is a constant. For this purpose we pass to the limit in the formula

$$P\{v = k\} = \binom{n}{k} P^k (1-P)^{n-k} =$$

$$= \frac{n(n-1)\ldots(n-k+1)}{k!} \left(\frac{\alpha}{n}\right)^k \left(1 - \frac{\alpha}{n}\right)^n$$

(where $\dot{P}\{v = k\}$ designates the probability that $v = k$) as $n \to \infty$. This results in the limiting formula

$$P\{v = k\} = \frac{\alpha^k}{k!} e^{-\alpha} \quad (k = 0, 1, 2, \ldots)$$

The calculations connected with the deduction of the formula are left to the reader. Thus, we arrive at a random variable which can assume infinitely many different values. The probability distribution thus obtained is illustrated by the following table:

values of v	0	1	2	...	k	...
probabilities	$e^{-\alpha}$	$\frac{\alpha}{1!} e^{-\alpha}$	$\frac{\alpha^2}{2!} e^{-\alpha}$	...	$\frac{\alpha^k}{k!} e^{-\alpha}$	...

This is the so-called **Poisson law** (**Poisson distribution**) named after S. Poisson (1781-1840), a French mechanician, physicist and mathematician.

An example of a random variable distributed according to the Poisson law is the number of atoms of a certain mass of some slowly disintegrating radioactive substance which decay during a sufficiently long time interval so chosen that it should be possible to observe the disintegrations of separate atoms. Then the Poisson law is in fact applicable because under these conditions the decay of

an atom is independent of the disintegrations of other atoms and all the atoms disintegrate with equal probability. There is a number of other similar examples.

10. Examples of Continuous Random Variables. One of the simplest examples is a random variable which is *uniformly distributed over an interval* $a \leqslant x \leqslant b$, that is which can assume all the values belonging to the interval with equal probability and does not assume the values lying outside the interval. The probability density of such a variable is put down in the form

$$p(x) = \begin{cases} c \text{ for } a \leqslant x \leqslant b \\ 0 \text{ for } x < a \text{ and for } x > b \end{cases}$$

Condition (11) implies that $c = \dfrac{1}{b-a}$. The graph of this function is shown in Fig. 343. A uniformly distributed random variable is sometimes said to have a **rectangular distribution** (Fig. 343 illustrates the origin of the name). For instance, the round-off error which
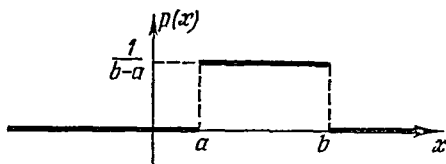
Fig. 343

results from rounding the numerical value of a quantity to its nearest integer is a uniformly distributed random variable, and we have $a = -0.5$, $b = 0.5$ and $c = 1$ in this case (why is it so?).

The most widely spread random variables are distributed according to the so-called **normal law (Gaussian law)**. The density function of such a random variable is expressed by the formula

$$p(x) = Me^{-\beta(x-\alpha)^2} = M \exp[-\beta(x-\alpha)^2]$$

where α, $M > 0$ and $\beta > 0$ are some numerical parameters. The parameter M can be easily expressed in terms of β. To achieve this we must take formula (11) and substitute $s = \sqrt{\beta}\,(x - \alpha)$ in it. Then, using integral (XIV.72), we deduce $M = \sqrt{\dfrac{\beta}{\pi}}$ (let the reader verify the calculations!). For our further aims (see Sec. 15) it will be convenient to introduce the notation $\beta = \dfrac{1}{2\sigma^2}$ and to put down the expression of $p(x)$ in the form:

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\alpha)^2}{2\sigma^2}\right] \tag{12}$$

The graph of density function (12) is shown in Fig. 344. In § 4 we shall discuss why the Gaussian law is so widely applied.

The probability that a random variable ξ distributed in accord with law (12) falls in an interval $a \leqslant x \leqslant b$ is equal to

$$P\{a \leqslant \xi \leqslant b\} = \frac{1}{\sqrt{2\pi}\,\sigma} \times$$

$$\times \int_a^b \exp\left[-\frac{(x-\alpha)^2}{2\sigma^2}\right] dx \quad (13)$$
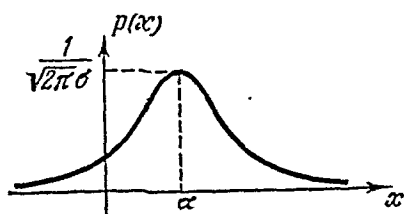
Fig. 344

The above probability can be easily expressed by means of the probability integral

$$\Phi(t) = \sqrt{\frac{2}{\pi}} \int_0^t \exp\left(-\frac{s^2}{2}\right) ds$$

for which there are extensive tables (for instance, see [23], [41] and [48]. Indeed, substituting $s = \frac{(x-\alpha)}{\sigma}$ into (13) we obtain the expression

$$P\{a \leqslant \xi \leqslant b\} = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\alpha}{\sigma}}^{\frac{b-\alpha}{\sigma}} \exp\left(-\frac{s^2}{2}\right) ds =$$

$$= \frac{1}{2}\sqrt{\frac{2}{\pi}}\left(\int_0^{\frac{b-\alpha}{\sigma}} \exp\left(-\frac{s^2}{2}\right) ds - \int_0^{\frac{a-\alpha}{\sigma}} \exp\left(-\frac{s^2}{2}\right) ds\right) =$$

$$= \frac{1}{2}\left[\Phi\left(\frac{b-\alpha}{\sigma}\right) - \Phi\left(\frac{a-\alpha}{\sigma}\right)\right] \quad (14)$$

11. Joint Distribution of Several Random Variables. We shall confine ourselves to the case of continuous random variables. Moreover, for simplicity's sake, we shall consider a system of two variables. Discrete variables and systems of more than two variables are investigated in a similar way.

Let us simultaneously consider two random variables ξ and η which take on certain numerical values in one and the same trial. Then the probability of ξ falling in an interval between x and $x + dx$ and η falling between y and $y + dy$ should be proportional both to dx and to dy. Hence, this probability is equal to an expression of the form $p(x, y)\, dx\, dy$. The function $p(x, y)$ is referred to as the probability density (frequency function) of the joint (simultaneous)

distribution of the random variables ξ and η. This function comple-tely characterizes the pair of random variables ξ, η. Obviously, the function must satisfy the conditions

$$p\,(x,\,y)\geqslant 0 \quad \text{and} \quad \int dx \int p\,(x,\,y)\,dy = 1$$

If the probability density of the joint distribution of two random variables ξ and η is known we can easily find the probability den-sity of each of the variables ξ and η (the densities of the so-called **marginal distributions** of ξ and η). Actually, formula (5) implies that the probability that ξ will assume a value lying between x and $x + dx$ when η can have an arbitrary value is equal to

$$\mathbf{P}\,\{x \leqslant \xi \leqslant x + dx\} = \int\limits_{y=-\infty}^{\infty} p\,(x,\,y)\,dx\,dy = \left(\int p\,(x,\,y)\,dy\right)dx$$

It follows, by Sec. 9, that the density of the probability distribu-tion of the variable ξ is a function $p_\xi\,(x)$ of the form

$$p_\xi\,(x) = \int p\,(x,\,y)\,dy$$

We similarly deduce the expression

$$p_\eta\,(y) = \int p\,(x,\,y)\,dx$$

for the probability density of η. But the converse transition from $p_\xi\,(x)$ and $p_\eta\,(y)$ to $p\,(x,\,y)$ may be impossible in the general case, that is it may be impossible to restore $p\,(x,\,y)$ knowing only $p_\xi\,(x)$ and $p_\eta\,(y)$, because here an essential role is played by the "interac-tion" between the variables ξ and η.

There is an important special case when $p\,(x,\,y)$ can be obtained on the basis of $p_\xi\,(x)$ and $p_\eta\,(y)$. This is the case when the random variables ξ and η are **independent**, i.e. when any information con-cerning one of them does not affect the probability of a numerical value assumed by the other. In this case formula (7) implies that

$$p\,(x,\,y)\,dx\,dy = \mathbf{P}\,\{x \leqslant \xi \leqslant x + dx,\;y \leqslant \eta \leqslant y + dy\} =$$
$$= \mathbf{P}\,\{x \leqslant \xi \leqslant x + dx\}\,\mathbf{P}\,\{y \leqslant \eta \leqslant y + dy\} = p_\xi\,(x)\,dx\cdot p_\eta\,(y)\,dy$$

i.e.

$$p\,(x,\,y) = p_\xi\,(x)\,p_\eta\,(y) \tag{15}$$

Conversely, if condition (15) holds we can show that the random variables ξ and η are independent.

There is a more general notion of a **multidimensional random variable** related to systems of random variables. Such a variable takes on the values which are the elements of a multidimensional

space (R) (see Sec. X.2). The law of probability distribution of a random variable of this type is represented by a non-negative measure defined in (R) (see Sec. XVI.19), the measure of the whole space (R) being equal to unity. The measure of a region belonging to the space is nothing but the probability that the variable in question falls in the region. If this measure in (R) is such that it is possible to perform differentiation with respect to it (for instance, such as the Lebesgue measure in a finite-dimensional Euclidean space mentioned in Sec. XVI.19) then, differentiating, we can pass to the probability density (see Sec. XVI.7). If we introduce generalized coordinates t_1, t_2, . . ., t_k in (R), we can consider the set of the coordinates of the element of the space (R) which represents the multi-dimensional random quantity in question instead of the quantity itself. Thus we come to a system of several random variables having a probability density of their joint distribution.

As a simple example illustrating what has just been said, we consider the n-dimensional normal (Gaussian) law. This is the law of probability distribution of a **random vector** ξ in the space E_n (see Sec. VII.18) whose probability density is of the form

$$p\,(\mathbf{x}) = M \exp\,(-\mathbf{x}^*A\mathbf{x}) \tag{16}$$

where A is a positive-definite symmetric matrix (see Secs. XI.11 and XII.7) and M is a **normalization factor** so chosen that the integral of $p\,(\mathbf{x})$ taken over the whole space should be equal to unity. As is known from Sec. XI.11, the quadratic form $\mathbf{x}^*A\mathbf{x}$ can be reduced to a diagonal form by means of introducing a new Cartesian basis in E_n. Hence, after the basis has been introduced, frequency function (16) is transformed to the form

$$p\,(\mathbf{x}') = M \exp\,(-\lambda_1 x_1'^2 - \lambda_2 x_2'^2 - \ldots - \lambda_n x_n'^2) =$$
$$= M \exp\,(-\lambda_1 x_1'^2) \exp\,(-\lambda_2 x_2'^2) \ldots \exp\,(-\lambda_n x_n'^2)$$

where λ_1, λ_2, . . ., λ_n are the eigenvalues of the matrix A. This enables us to easily find $M = \sqrt{\lambda_1 \lambda_2 \ldots \lambda_n}\, \pi^{-n/2}$. Besides, formula (15) indicates that the coordinates of the random vector with respect to the new basis are independent random variables.

12. **Functions of Random Variables.** If $\eta = f\,(\xi)$ where ξ is a random variable, η is also a random variable. Besides, if ξ is discrete (continuous), η is also discrete (continuous). If ξ is represented by a table of form (10) then, generally speaking, η will assume the value $f\,(x_1)$ with the probability p_1, the value $f\,(x_2)$ with the probability p_2 and so on. But at the same time we must take into account the fact that if $f\,(x_i) = f\,(x_j)$ (where $i \neq j$) then, of course, the corresponding probabilities p_i and p_j are added together. For instance, if ξ takes on the values -2, -1, 0, 1 and 2 with the same probabi-

lity $\frac{1}{5}$, it follows that ξ^2 takes on the values 0, 1 and 4 with the corresponding probabilities $\frac{1}{5}, \frac{2}{5}$ and $\frac{2}{5}$ (why?).

If $\eta = f(\xi)$ and ξ is a *continuous* random variable with the probability density $p_\xi(x)$ then in the case $\eta = f(\xi)$ is an increasing function of ξ its probability density $p_\eta(y)$ is expressed by the formula

$$p_\eta(y) = \frac{1}{dy} \, \mathbf{P}\{y \leqslant \eta \leqslant y + dy\} = \frac{1}{dy} \, \mathbf{P}\{x \leqslant \xi \leqslant x + dx\} =$$

$$= \frac{1}{dy} \, p_\xi(x) \, dx = \frac{p_\xi(x)}{f'(x)}$$

where x entering into the right-hand side is found from the equation $f(x) = y$. If the function $f(\xi)$ is a decreasing one $|f'(x)|$ should be substituted for $f'(x)$ into the right-hand side. Finally, if the function $f(\xi)$ is a non-monotone one, the right-hand side should be replaced by the sum of analogous expressions, the summation being extended over all the solutions of the equation $f(x) = y$.

We can similarly investigate functions of several random variables. For example, let us take a function of the form $\zeta = f(\xi, \eta)$ where the pair of random variables ξ, η is characterized by the density of their joint probability distribution $p(x, y)$. Then

$$p_\zeta(z) = \frac{1}{dz} \, \mathbf{P}\{z \leqslant \zeta \leqslant z + dz\} = \frac{d}{dz} \iint\limits_{f(x,\,y) \leqslant z} p(x, y) \, dx \, dy \qquad (17)$$

In the general case the derivative entering into the right-hand side of (17) is computed according to the rules given in Sec. XVI.18.

We shall illustrate the above result by applying it to calculating the probability density $p_\zeta(z)$ of the sum $\zeta = \xi + \eta$ of two independent random variables ξ and η. By formulas (15) and (17), we obtain

$$p_\zeta(z) = \frac{d}{dz} \iint\limits_{x+y \leqslant z} p_\xi(x) \, p_\eta(y) \, dx \, dy =$$

$$= \frac{d}{dz} \int dx \int\limits_{-\infty}^{z-x} p_\xi(x) \, p_\eta(y) \, dy = \frac{d}{dz} \int \left[\int\limits_{-\infty}^{z-x} p_\eta(y) \, dy \right] p_\xi(x) \, dx =$$

$$= \int \left[\frac{d}{dz} \int\limits_{-\infty}^{z-x} p_\eta(y) \, dy \right] p_\xi(x) \, dx = \int p_\xi(x) \, p_\eta(z-x) \, dx$$

(let the reader verify the calculations!).

As an exercise, we suggest that the reader should prove that the sum of two independent random variables each of which is uniform-

ly distributed over the same interval $0 \leqslant x \leqslant 1$ has the probability density of the form

$$p(x) = \begin{cases} 0 & \text{for } x < 0 \text{ and } x > 2 \\ x & \text{for } 0 \leqslant x \leqslant 1 \\ 2 - x & \text{for } 1 < x \leqslant 2 \end{cases}$$

§ 3. Numerical Characteristics of Random Variables

13. The Mean Value. Let there be a discrete random variable ξ represented by a table of form (10). Suppose that a great number N of trials have been performed. What will be the arithmetic mean of the values ξ thus obtained? To answer the question we denote by N_i the number of the outcomes of the trials in which ξ has assumed the value x_i. Then the sought-for arithmetic mean is equal to

$$\frac{N_1 x_1 + N_2 x_2 + N_3 x_3 + \cdots}{N} = x_1 \frac{N_1}{N} + x_2 \frac{N_2}{N} + x_3 \frac{N_3}{N} + \cdots$$

But as we know from Sec. 2, we have $\frac{N_i}{N} \to P_i$ when $N \to \infty$. Hence, in the limit, we obtain the expression

$$x_1 P_1 + x_2 P_2 + x_3 P_3 + \cdots \tag{18}$$

It is referred to as the **mean value (mathematical expectation or, briefly, expectation or centre of distribution)** of the random variable ξ. This is one of the most important characteristics of ξ. The mean of ξ is usually designated as $\overline{\xi}$ or $M\{\xi\}$ or $M\xi$. It should be noted that the mean value of a random variable is no longer a random variable but is a deterministic quantity. (For instance, verify that the mean value of the sums of points obtained in throwing a die is the constant number 3.5.)

Formula (18) can be obviously generalized for the case when ξ is a continuous random variable with the probability density $p(x)$:

$$\overline{\xi} = \sum x_i dP = \sum x p(x) \, dx = \int x p(x) \, dx \tag{19}$$

[We have put down the summation sign to stress the analogy between formulas (18) and (19); of course there must be the sign of integration here which has been written in the last expression entering into (19).] Let the reader verify, by means of the formula, that the means of the variables considered in Sec. 10 are, respectively, $\frac{a+b}{2}$ and α. But these results are obviously implied by the symmetry of the distributions.

14. Properties of the Mean Value.

1. The definition implies that the mean value $\bar{\xi}$ of a random variable ξ has the same dimension as ξ and that it lies between the greatest and the least possible values of ξ.

2. If we multiply a random variable by a constant (i.e. by a constant deterministic quantity) its mean value will be multiplied by the same constant: $M\{C\xi\} = CM\{\xi\}$ ($C = $ const). This follows from Sec. 13 because the multiplication of all the values by a constant yields the multiplication of the arithmetic mean value by the same constant. The next property is proved in a similar way.

3. The mean of the sum of two random variables equals the sum of their means $M\{\xi + \eta\} = M\{\xi\} + M\{\eta\}$. In particular, if a constant is added to a random variable, the same constant is added to its mean value.

By the way, applying the last property we can readily find the mean value of a random variable ξ distributed according to the binomial law (see Sec. 9). Let us consider independent random variables ξ_1, ξ_2, ..., ξ_n which take on the value 1 with the probability P and the value 0 with the probability $1 - P$. Apparently, we can interpret ξ_i as a variable indicating the number of the occurrences of an event A in the ith trial, the probability of A being P. Then the variable in question can be represented in the form

$$\xi = \xi_1 + \xi_2 + \ldots + \xi_n \tag{20}$$

(see Sec. 9 where an analogous variable was denoted by ν). From (20) we obtain $M\{\xi\} = M\{\xi_1\} + M\{\xi_2\} + \ldots + M\{\xi_n\} = nP$. This result is directly implied by the meaning of the variable, and we could have guessed it without applying the above calculations. It follows that for a Poisson distribution (Sec. 9) we have $M\{\xi\} = \alpha$.

4. The mean of the product of two independent random variables is equal to the product of their mean values:

$$M\{\xi\eta\} = M\{\xi\} M\{\eta\} \text{ if } \xi \text{ and } \eta \text{ are independent}$$

Indeed, if ξ takes the values x_i with the probabilities p_i and η takes the values y_j with the probabilities q_j then $\xi\eta$ assumes the values $x_i y_j$ with the probabilities $p_i q_j$ since ξ and η are independent (see Sec. 4). Therefore

$$M\{\xi\eta\} = \sum_{i,j} x_i y_j (p_i q_j) = \sum_i \left(\sum_j x_i y_j p_i q_j \right) =$$
$$= \sum_i x_i p_i \sum_j y_j q_j = M\{\xi\} M\{\eta\}$$

Properties 3 and 4 are immediately extended to an arbitrary number of summands and factors. It should be noted that the condition that the factors should be independent is essential for property 4. If the condition does not hold the property no longer remains true in the general case. For example, if we square a random variable,

i.e. multiply it by itself, then, as a rule, the mean value of the square does not equal the square of the mean: $\overline{\xi^2} \neq (\overline{\xi})^2$. For instance, in the example considered at the beginning of Sec. 12 we have $M\{\xi\} = 0$ but $M\{\xi^2\} = 2$ (check it up!).

5. If a random variable ξ assumes the values which are placed symmetrically with respect to a constant a with equal probabilities, then $\overline{\xi} = a$ (this is obvious).

6. If a random variable ξ is represented by a table of form (10) it follows that $\overline{f(\xi)} = \sum_i f(x_i)\, p_i$. If a continuous random variable

ξ has the probability density $p(x)$ we have $\overline{f(\xi)} = \int f(x)\, p(x)\, dx$ (this immediately follows from the definitions).

In particular, the calculation of the mean of a random variable enables us to set a criterion according to which a random event can be considered to be practi-
cally impossible (that is to set a certain value of the quanti-
ty ε mentioned in Sec. 7). Here we shall give only some simple considerations concerning this question. Suppose we have agreed that the random events whose probabilities are less than a cer-
tain value ε are disregarded, i.e. they are considered to be practi-
cally impossible. But there can be an incorrect prediction, and this means that an event which is regarded as impossible may nevertheless occur. Let the loss connected with the incorrect pre-
diction be equal to an amount k expressed in certain monetary units. Then the average (mean) loss will be equal to εk. It is therefore desirable to decrease ε,

Fig. 345

but at the same time the perfection of the predictions also involves some additional expenditure. Let us designate the cost of a trial which can guarantee a prediction "accurate to ε" as $f(\varepsilon)$. In many concrete problems such a function can be approximately found. An example of the graph of a function of this type is represented in Fig. 345. Hence, the average loss connected with incorrect predic-
tions equals $f(\varepsilon) + k\varepsilon$, and thus the value $\varepsilon = \varepsilon_0$ which is to be set as a criterion must be chosen so that this sum should be mini-
mized.
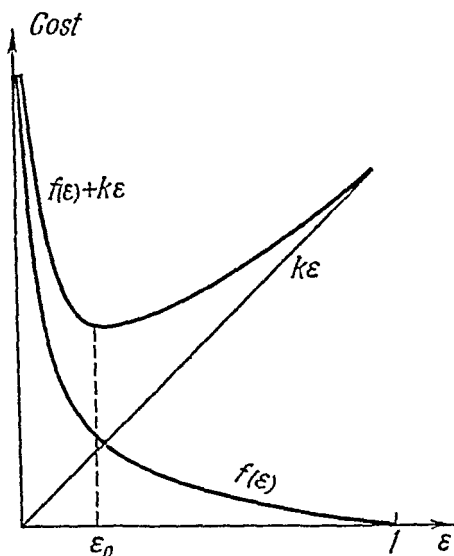
In conclusion we shall give several remarks concerning multidimensional random quantities (see Sec. 11) which assume their values in a finite-dimensional linear space (R) (see Sec. VII.17). Thus, we shall speak about finite-dimensional random vectors. The formula of the mean value of such a random variable is similar to (19):

$$\bar{\xi} = \int\limits_{(R)} x\, dP = \int\limits_{(R)} x p(x)\, dR$$

where the integration is extended over the whole space (R) and dP is the differential of volume (the element of volume) in (R). All the properties of the mean in this case are analogous to those of the scalar (one-dimensional) case. Besides, property 4 holds for all kinds of products in which it is permissible to remove brackets according to the ordinary arithmetical rules (e.g. for the product of a vector by a scalar, for the scalar or vector product of vectors and so on).

15. Variance. The variance characterizes the degree of the spread of a random variable about its mean (expectation). Let us be given a random variable ξ. By definition, its **variance** (also called **dispersion**) is the quantity

$$D\xi = D\{\xi\} = M\{(\xi - M\xi)^2\} \tag{21}$$

This quantity is deterministic and always positive except the case when ξ itself is a deterministic quantity (in this case we have $D\xi = 0$).

By property 6 in Sec. 14, formula (21) implies the formulas

$$D\xi = \sum_i (x_i - \bar{\xi})^2 P_i \quad \text{and} \quad D\xi = \int (x - \bar{\xi})^2 p(x)\, dx$$

From (21), we easily see that if ξ is multiplied by a constant C then $D\xi$ is multiplied by C^2 and that if a constant is added to ξ its variance $D\xi$ does not change. Further, if two random variables ξ and η are independent, we have

$$D\{\xi + \eta\} = D\xi + D\eta \tag{22}$$

In fact,

$$D\{\xi + \eta\} = M\{[\xi + \eta - M(\xi + \eta)]^2\} = M\{[(\xi - M\xi) +$$
$$+ (\eta - M\eta)]^2\} = M(\xi - M\xi)^2 + 2M\{(\xi - M\xi)(\eta - M\eta)\} +$$
$$+ M(\eta - M\eta)^2 = D\xi + 2M(\xi - M\xi)\cdot M(\eta - M\eta) + D\eta =$$
$$= D\xi + 2\cdot 0 \cdot 0 + D\eta = D\xi + D\eta$$

(where has the independence of the random variables ξ and η been used in the above calculations?).

Let us determine the dispersions for the examples considered in Secs. 9 and 10. The variance of each summand entering into formula (20) is equal to $(0 - P)^2 (1 - P) + (1 - P)^2 P = P (1 - P)$. Hence, by formula (22), we obtain the expression $D\xi = nP (1 - P)$ for the binomial law. Now passing to the limit, as $n \to \infty$, we obtain the expression $D\xi = \alpha$ for the Poisson distribution. For the uniform distribution over an interval $a \leqslant x \leqslant b$ we deduce

$$D\xi = \int \left(x - \frac{a+b}{b} \right)^2 p(x)\, dx = \int_a^b \left(x - \frac{a+b}{2} \right)^2 \frac{1}{b-a}\, dx = \frac{(b-a)^2}{12}$$

(check up the result!). Finally, for the normal law we obtain

$$D\xi = \int (x-\alpha)^2 \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\alpha)^2}{2\sigma^2} \right] dx$$

Substituting $s = \dfrac{x-\alpha}{\sqrt{2}\,\sigma}$ into the integral we get

$$D\xi = \frac{2\sigma^2}{\sqrt{\pi}} \int s^2 \exp\left(-s^2 \right) ds$$

Now, putting $s = u$ and $dv = s \exp(-s^2)$ $\Big[$that is $ds = du$ and $v = -\dfrac{1}{2}\exp(-s^2)\Big]$ we integrate by parts and thus deduce the result

$$D\xi = \frac{\sigma^2}{\sqrt{\pi}} \int \exp\left(-s^2 \right) ds = \sigma^2$$

Together with the variance $D\xi$, we often use the square root of it which is called the **standard deviation** of the random variable ξ. The standard deviation $\sqrt{D\xi}$ is of the same dimension as ξ. We see that the parameter σ entering into Gaussian law (12) is nothing but the standard deviation of the normally distributed random variable ξ.

Formula (22) implies an important consequence. Let random variables $\xi_1, \xi_2, \ldots, \xi_n$ be independent and let them be distributed according to the same law with the standard deviation σ. Then their sum has the dispersion $n\sigma^2$, and therefore its standard deviation is equal to $\sqrt{n}\sigma$. Now notice that the above two examples indicate that the values of a random variable which correspond to the higher probability are concentrated on an interval whose length is directly proportional to the standard deviation (this property will be discussed in more detail in Sec. 19). Hence, for the sum of independent random summands, the length of such an interval is proportional to $\sqrt{n}$ (but not to n as it would be if we had n equal summands). In particular, this law holds for the error of the sum of several summands which are known with the same accuracy (compare with Sec. I.9).

Two different random variables distributed according to different laws may have the same expectations and the same variances. Therefore, to obtain a more complete description of random variables, we also use some other numerical characteristics. In particular, we introduce the so-called **moments** of a random variable (of its probability distribution) which are defined as

$$M\{\xi^k\} = \int x^k p(x)\,dx \quad (k = 1, 2, 3, \ldots)$$

under the assumption that the corresponding integrals are convergent. This is the kth moment (the moment of order k) of the variable ξ. The first moment is nothing but the expectation (mean value), and the variance, as it is implied by formula (21), is expressed in terms of the moments of the second order:

$$D\xi = M\{\xi^2\} - 2M\{\xi M\xi\} + M\{(M\xi)^2\} = M\{\xi^2\} - (M\xi)^2$$

The higher-order moments characterize the law of probability distribution of a random variable more completely than the expectation and the variance.

16. **Correlation.** Let us be given the probability density $p(x, y)$ of a joint distribution of two random variables ξ and η (see Sec. 11). If it is known that the variable ξ has assumed the value $\xi = a$ we can speak about the **conditional probability distribution** of the random variable η *relative to the hypothesis that ξ takes the value a.* Let us denote the corresponding **conditional density function** of η *relative to the hypothesis* $\xi = a$ as $p_\eta(y \mid \xi = a)$*. Then the conditional probability of η falling in an interval between y and $y + dy$ provided ξ takes the value $\xi = a$ is equal to $p_\eta(y \mid \xi = a)\,dy$, and

$$\int p_\eta(y \mid \xi = a)\,dy = 1$$

---

* *Translator's note.* As it was shown in Sec. 11, knowing the density $p(x, y)$ of the joint distribution we can find the densities of the marginal distributions $p_\xi(x) = \int p(x, y)\,dy$ and $p_\eta(y) = \int p(x, y)\,dx$ of the random variables ξ and η. Now, according to formula (6), we can write

$$P(y \leqslant \eta \leqslant y + dy \mid \xi = a) = \lim_{dx \to 0} \frac{P(a \leqslant \xi \leqslant a + dx,\ y \leqslant \eta \leqslant y + dy)}{P(a \leqslant \xi \leqslant a + dx)} =$$

$$= \lim_{dx \to 0} \frac{p(a, y)\,dx\,dy}{p_\xi(a)\,dx} = \frac{p(a, y)}{p_\xi(a)}\,dy$$

which implies

$$p_\eta(y \mid \xi = a) = \frac{p(a, y)}{p_\xi(a)}$$

The conditional probability density $p_\xi(x \mid \eta = b)$ of ξ relative to the hypothesis $\eta = b$ is found in like manner:

$$p_\xi(x \mid \eta = b) = \frac{p(x, b)}{p_\eta(b)}$$

When the value of ξ entering into the hypothesis varies the law of the conditional probability distribution of the variable η changes in the general case. Hence, there is a certain relationship between ξ and η but it differs from an ordinary functional relationship between deterministic variables which was studied in the foregoing chapters.

A relationship of this kind is called a **correlation**. Similarly, if we are given a joint distribution of an arbitrary number of random variables we can define the correlation between any of the variables and the rest.

. We often encounter relationships of a correlation type. For instance, when we speak about the relationship between the weight of a person and his height we undoubtedly mean a correlative relation because we know that the weight is not completely and uniquely specified by the height. At the same time it is quite clear that the law of distribution of the weights of the people two metres high differs from that of the people one and a half metres high. When we say that smoking reduces the duration of life of a person we also mean a correlative dependence because, although there are many cases of different kind, we nevertheless find that the average duration of life of non-smokers is higher than that of smokers if we consider the law of probability distribution of the duration of life. We must carefully distinguish between the deterministic and correlative dependences and also take into account that in the latter case the existence of contradictory examples does not affect the general validity of probability inferences.

The mean value of the conditional probability distribution of the random variable η is a deterministic function of x (if x designates the numerical value assumed by the variable ξ which we denoted as $\xi = x = a$ above):

$$M\{\eta \,|\, \xi = x\} = \int y p_\eta \,(y\,|\,\xi = x)\,dy$$

Let us denote this function as $f(x)$. The function $f(x)$ (called the **regression function of η on ξ**) expresses the **conditional mean value** of η relative to the hypothesis $\xi = x$; the graph of the function is referred to as the **regression curve for the mean of η** (the **regression line of η on ξ**). In the above example of the duration of life of smokers, it is the regression that defines the regularities we are interested in. We can similarly determine the conditional mean value $\varphi(y)$ of ξ relative to the hypothesis that $\eta = y$ and construct the corresponding regression curve for the mean of ξ. It is interesting that, generally speaking, the functions $f(x)$ and $\varphi(y)$ are not inverse with respect to each other as it would be if we had a deterministic relationship. This becomes especially clear in the case of independent random variables ξ and η when both conditional means are con-

stant and the corresponding regression curves turn into straight lines parallel to the y-axis and to the x-axis, respectively.

There is a comparatively simple special case when the regression of both variables ξ and η is linear, i.e. when both regression functions $f(x)$ and $\varphi(y)$ are linear. Let us introduce the notation

$$r_{\xi,\,\eta} = \frac{\overline{\xi\eta} - \overline{\xi}\overline{\eta}}{\sqrt{D\xi D\eta}}$$

The quantity $r_{\xi,\,\eta}$ is called the **correlation coefficient** of the random variables ξ and η. It is possible to prove that we always have $|\,r_{\xi,\,\eta}\,| \leqslant$ $\leqslant 1$ and that in the case when both functions $f(x)$ and $\varphi(y)$ are linear they have the form

$$f(x) = r_{\xi,\,\eta}\sqrt{\frac{D\eta}{D\xi}}\,(x - \overline{\xi}) + \overline{\eta} \quad \text{and} \quad \varphi(y) = r_{\xi,\,\eta}\sqrt{\frac{D\xi}{D\eta}}\,(y - \overline{\eta}) + \overline{\xi}$$

but we shall not give the proof here. It follows that if $r_{\xi,\,\eta} > 0$ both functions are increasing and if $r_{\xi,\,\eta} < 0$ they are decreasing.

An important example of a linear correlation is the two-dimensional normal law (see Sec. 11) with the density function

$$p(x,\,y) = M \exp\left[-(Ax^2 + 2Bxy + Cy^2)\right]$$

where M is the normalization factor and the quadratic form in the parentheses is positive-definite. We suggest that the reader prove that in this case we have

$$f(x) = -\frac{B}{C}\,x, \quad \varphi(y) = -\frac{B}{A}\,y \quad \text{and} \quad r_{\xi,\,\eta} = -\frac{B}{\sqrt{AC}}$$

17. Characteristic Functions. The *characteristic function* of a random variable ξ is a function of a real parameter u of the form

$$\varphi_{\xi}(u) = M\{e^{iu\xi}\} \quad (-\infty < u < \infty)$$

Property 6 in Sec. 14 enables us to write in full the expression of a characteristic function in the form of a sum or of an integral:

$$\varphi_{\xi}(u) = \sum_{h} P_{h}e^{iux_{h}} \quad \text{or} \quad \varphi_{\xi}(u) = \int e^{iux}p_{\xi}(x)\,dx \qquad (23)$$

For the first time characteristic functions were systematically employed by A. M. Lyapunov.

The second formula (23) is nothing but the Fourier integral of the function $\varphi_{\xi}(u)$ [see formula (XVII.141) in which we used another notation]. Hence, the probability density $p_{\xi}(x)$ is expressed in terms of the characteristic function by the formula

$$p_{\xi}(x) = \frac{1}{2\pi}\int \varphi_{\xi}(u)\,e^{-iux}\,du$$

We now enumerate some simple properties of a characteristic function. Formula (23) shows that there must always be $|\varphi_\xi(u)| \leqslant 1$ and $\varphi_\xi(0) = 1$. If $\eta = C_1\xi + C_2$ (where C_1 and C_2 are constants) then

$$\varphi_\eta(u) = M\{e^{iu(C_1\xi+C_2)}\} = M\{e^{iC_2u}e^{iC_1u\xi}\} = e^{iC_2u}\varphi_\xi(C_1u)$$

If $\zeta = \xi + \eta$ and the variables ξ and η are independent then

$$\varphi_\zeta(u) = M\{e^{iu(\xi+\eta)}\} = M\{e^{iu\xi}e^{iu\eta}\} = M\{e^{iu\xi}\}\,M\{e^{iu\eta}\} = \varphi_\xi(u)\,\varphi_\eta(u)$$

The first formula (23) and the last property enable us to deduce the expression for the characteristic function of a random variable having a binomial distribution (see Sec. 9):

$$\varphi_\xi(u) = (1 - P + Pe^{iu})^n$$

Now passing to the limit we obtain the characteristic function of a random variable distributed according to the Poisson law:

$$\varphi_\xi(u) = \exp(-\alpha + \alpha e^{iu})$$

For the case of a uniform distribution (see Sec. 10) we obtain

$$\varphi_\xi(u) = \frac{(e^{ibu} - e^{iau})}{i(b-a)u}$$

For our further aims it is necessary to find the Fourier transform of the function $f(x) = \exp(-x^2)$. Applying formula (XVII.138) we obtain

$$\hat{f}(k) = \frac{1}{2\pi}\int \exp(-x^2 - ikx)\,dx =$$

$$= \frac{1}{2\pi}\exp\left(-\frac{k^2}{4}\right)\int \exp\left[-\left(x + i\frac{k}{2}\right)^2\right]dx$$

(Check it up!) But the last integral in fact does not depend on k. Indeed, denoting the integral by $I(k)$ and differentiating we obtain

$$\frac{dI}{dk} = -\int \exp\left[-\left(x + i\frac{k}{2}\right)^2\right]2\left(x + i\frac{k}{2}\right)\frac{i}{2}\,dx =$$

$$= \frac{i}{2}\exp\left[-\left(x + i\frac{k}{2}\right)^2\right]\Big|_{x=-\infty}^{x=\infty} = 0$$

(why is it so?). Consequently, $I(k) = I(0) = \int \exp(-x^2)\,dx = \sqrt{\pi}$. Thus, finally we get $\hat{f}(k) = \frac{1}{2\sqrt{\pi}}\exp\left(-\frac{k^2}{4}\right)$. From this, with the help of property 3 in Sec. XVII.33, we conclude that the Fourier transform of the function $\exp(-ax^2)$ (where $a > 0$) is the function $\frac{1}{2\sqrt{\pi a}}\exp\left(-\frac{k^2}{4a}\right)$.

Now we can readily determine the characteristic function of a random variable distributed according to the Gaussian law (see Sec. 10). Let us first take the case $\alpha = 0$. Formula (23) expressing the Fourier inverse transform of the function $p_\xi(x)$, it is sufficient to multiply our result by 2π which yields

$$\varphi_\xi(u) = 2\pi \frac{1}{\sqrt{2\pi}\,\sigma} \frac{1}{2} \sqrt{\frac{2\sigma^2}{\pi}} \exp\left(-\frac{u^2}{4}\,2\sigma^2\right) = \exp\left(-\frac{\sigma^2 u^2}{2}\right)$$

To investigate the general case when $\alpha \neq 0$ we can add α to the above random variable for which the characteristic function has just been computed. Then denoting the new variable by the same letter ξ and taking advantage of the properties of characteristic functions enumerated above we finally obtain

$$\varphi_\xi(u) = e^{i\alpha u} \exp\left(-\frac{\sigma^2 u^2}{2}\right) = \exp\left(i\alpha u - \frac{\sigma^2 u^2}{2}\right)$$

In particular, this result implies a remarkable consequence. Let ξ_1 and ξ_2 be two independent random variables distributed according to the normal law with the parameters α_1, σ_1 and α_2, σ_2, respectively. Then, for the variable $\xi = \xi_1 + \xi_2$, we obtain

$$\varphi_\xi(u) = \varphi_{\xi_1}(u)\,\varphi_{\xi_2}(u) = \exp\left(i\alpha_1 u - \frac{\sigma_1^2 u^2}{2}\right) \exp\left(i\alpha_2 u - \frac{\sigma_2^2 u^2}{2}\right) =$$
$$= \exp\left[i\,(\alpha_1 + \alpha_2)\,u - \frac{(\sigma_1^2 + \sigma_2^2)\,u^2}{2}\right]$$

Thus, we have again arrived at the normal law with the parameters $\alpha = \alpha_1 + \alpha_2$ and $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$. The invariance of the normal law with respect to the addition of random variables is one of the basic properties of the law which accounts for its being so widely spread. Among the probability distributions of discrete random variables, the Poisson law possesses this property.

§ 4. Applications of the Normal Law

18. The Normal Law as the Limiting One. We now investigate the behaviour of the binomial law (see Sec. 9) when P remains constant and $n \to \infty$. A random variable $\xi^{(n)}$ distributed according to the binomial law has the mean $a_n = nP$ and the standard deviation $\sigma_n = \sqrt{n}\,\sqrt{P\,(1 - P)}$ (see Secs. 14 and 15) and hence we have $a_n \to \infty$ and $\sigma_n \to \infty$ for $n \to \infty$. Thus, we see that $\xi^{(n)}$ "spreads" over the whole x-axis in the limit. This makes it difficult to investigate the behaviour of $\xi^{(n)}$ directly. It is therefore convenient to perform a linear transformation of the variable $\xi^{(n)}$ so that the mean value should become equal to zero and the standard deviation

become equal to unity after the transformation has been carried out. A transformation of this kind is called the **standardization** (or normalization) of the variable $\xi^{(n)}$, and it is expressed by the following simple formula:

$$\eta^{(n)} = \frac{1}{\sigma_n} (\xi^{(n)} - a_n)$$

The variable $\eta^{(n)}$ is known as the **standardized (normalized) variable** corresponding to the random variable $\xi^{(n)}$.

There is a remarkable theorem referred to as the **De Moivre-Laplace theorem** which states that the law of distribution of the above standardized random variable tends to the normal law when $n \to \infty$. (P. Laplace, 1749-1827, a famous French astronomer, physicist and mathematician.)

The theorem is proved as follows. By Sec. 17, we have

$$\varphi_{\eta^{(n)}}(u) = e^{-i\frac{a_n}{\sigma_n}u}\left(1 - P + Pe^{i\frac{u}{\sigma_n}}\right)^n =$$

$$= \exp\left(-i\sqrt{\frac{nP}{1-P}}\,u\right)\left[1 - P + P\exp\left(\frac{iu}{\sqrt{nP(1-P)}}\right)\right]^n =$$

$$= \left\{\exp\left(-i\sqrt{\frac{P}{n(1-P)}}\,u\right)\left[1 - P + P\exp\left(\frac{iu}{\sqrt{nP(1-P)}}\right)\right]\right\}^n$$

Expanding the expression in the curly brackets in powers of $\frac{1}{\sqrt{n}}$ we obtain

$$\varphi_{\eta^{(n)}}(u) = \left\{\left(1 - i\sqrt{\frac{P}{n(1-P)}}\,u - \frac{Pu^2}{2n(1-P)} + \cdots\right) \times\right.$$

$$\times\left.\left[1 - P + P\left(1 + \frac{iu}{\sqrt{nP(1-P)}} - \frac{u^2}{2nP(1-P)} + \cdots\right)\right]\right\}^n =$$

$$= \left\{1 - \frac{u^2}{2n} + \cdots\right\}^n \xrightarrow[n\to\infty]{} \exp\left(-\frac{u^2}{2}\right)$$

(check up the calculations!). Thus we have arrived at the characteristic function of the normal law (see Sec. 16) with the parameters $\alpha = 0$ and $\sigma = 1$.

In Sec. 14 we mentioned that a variable distributed according to the binomial law is the sum of n independent random summands with the same simplest law of probability distribution. But it turns out that the normal law is obtained in the limit for any initial law of distribution (of course, except a deterministic law).

Indeed, for the sake of simplicity, let us suppose that we have an initial law of distribution with the characteristic function $\varphi_0(u)$ and with the zero mean value, the variance being equal to unity. These limitations are inessential because in the general case we carry

out the normalization. Then formula (23) implies $\varphi_0'(0) = 0$ and $\varphi_0''(0) = -1$ etc., and hence, on the basis of Taylor's formula, we have $\varphi_0(u) = 1 - \frac{u^2}{2} + \ldots$. Using the notation similar to the above we find

$$\varphi_{\eta(n)}(u) = \left[\varphi_0\left(\frac{u}{\sqrt{n}}\right)\right]^n = \left[1 - \frac{u^2}{2n} + \ldots\right]^n \xrightarrow[n \to \infty]{} \exp\left(-\frac{u^2}{2}\right)$$

It turns out that the condition that the laws of distribution of the summands should be the same is also inessential. For instance, A. M. Lyapunov proved that the law of distribution for the standardized sum of independent random summands ξ_1, ξ_2, ..., ξ_n is also close to the Gaussian law when n is large if the ratio

$$\sum_{k=1}^n \mathbf{M}|\xi_k - a_k|^3 : \left(\sum_{k=1}^n \mathbf{D}\xi_k\right)^{\frac{3}{2}} \quad (a_k = \mathbf{M}\xi_k)$$

is small. This condition is violated if the variance of a small number of summands considerably exceeds the variance of the rest. In this case the latter summands do not contribute to the whole result, in the limit, after the standardization has been performed. Lyapunov's condition is also violated in some other special cases, for instance, in the case leading to the Poisson law (check up this assertion!).

If the standardization results in a normal distribution we obviously have a normally distributed variable before the standardization but with an arbitrary mean value and variance. Hence, we can say that the sum of many independent random summands is normally distributed irrespective of the laws of distribution of the summands. The exceptions to the rule are the cases enumerated in the foregoing paragraph. Here lies the main cause making the Gaussian law so important.

In particular, it is usually assumed that the random errors of a measurement obey the normal law. Actually, as a rule, such an error results from mutual superposition of a great many small independent errors which cannot be taken into account separately. It is this fact that leads to the assumption that the Gaussian law is applicable here.

19. Confidence Interval. We now come back to the problem of tossing a coin which was considered in Sec. 1. It is clear that if the coin comes up heads 200 times in 1000 tosses we have every reason to suspect that something is wrong. Shall we say the same if we have 400 heads or 450 heads? In other words, shall we consider it to be unusual if the relative frequency of the coin coming up heads is 0.4 or 0.45? Now we are able to answer a question of this type.

Let us take a more general situation. Consider a random variable ξ (in the above example the role of ξ was played by the number of occurrences of heads in one toss). Suppose that n trials have been

performed, and let ξ assume the values $x_1,\ x_2,\ \ldots,\ x_n$ in these trials. Let us designate the arithmetic mean value of the quantities $x_1,\ x_2,\ \ldots,\ x_n$ by $x^{(n)}$:

$$x^{(n)} = \frac{1}{n}\,(x_1 + x_2 + \ldots + x_n) \tag{24}$$

On the basis of Sec. 13, we can assert that $x^{(n)} \to \bar{\xi}$ for $n \to \infty$. This is the so-called law of large numbers (in our course we have taken the law of large numbers as the foundation of the definition of the mean $\bar{\xi}$ of a random quantity ξ). But what is the rate at which $x^{(n)}$ approaches $\bar{\xi}$, as $n \to \infty$?

Let us consider the random variable

$$\xi^{(n)} = \frac{1}{n}\,(\xi_1 + \xi_2 + \ldots + \xi_n)$$

where all the summands are independent, each ξ_i $(i = 1, 2, \ldots, n)$ being distributed according to the same law of probability distribution as ξ. Then quantity (24) is one of the possible values of the variable $\xi^{(n)}$. But, on the basis of Sec. 18, we can regard the random variable $\xi^{(n)}$ as being distributed according to the Gaussian law for large n. Besides, we have

$$\overline{\xi^{(n)}} = \frac{1}{n}\,(\bar{\xi}_1 + \bar{\xi}_2 + \ldots + \bar{\xi}_n) = \bar{\xi}$$

and

$$D\xi^{(n)} = \frac{1}{n^2}\,n D\xi = \frac{1}{n}\,D\xi = \frac{1}{n}\,\sigma_\xi^2$$

where σ_ξ is the standard deviation of the variable ξ. Therefore, by formula (14), we can put

$$P\{a \leqslant \xi^{(n)} \leqslant b\} = \frac{1}{2}\left[\Phi\left(\sqrt{n}\,\frac{b-\bar{\xi}}{\sigma_\xi}\right) - \Phi\left(\sqrt{n}\,\frac{a-\bar{\xi}}{\sigma_\xi}\right)\right]$$

For the sake of simplicity, let us restrict ourselves to the case of symmetric intervals of the form $|\xi^{(n)} - \bar{\xi}| \leqslant \delta$. Then we obtain

$$P\{|\xi^{(n)} - \bar{\xi}| \leqslant \delta\} = \Phi\left(\frac{\delta\sqrt{n}}{\sigma_\xi}\right) \tag{25}$$

It is formula (25) that provides the answer to the problem stated at the beginning of this section. In practical applications it can be used for any values of n. Here we give a rough table of the values of the function $\Phi(t)$:

t	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1
$\Phi(t)$	0.00	0.08	0.16	0.24	0.31	0.38	0.45	0.52	0.58	0.63	0.68	0.73

t	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4
$\Phi(t)$	0.77	0.81	0.84	0.87	0.89	0.91	0.93	0.94	0.95	0.96	0.97	0.98	0.98

t	2.6	3.0	3.3	3.9	4.4	4.9	5.3	∞
$\Phi(t)$	0.99	0.997	0.999	0.9999	0.99999	0.999999	0.9999999	1

For large values of t the corresponding values of $\Phi(t)$ can be found with a great accuracy by means of the asymptotically convergent series (see Sec. XVII.19):

$$\Phi(t) = \frac{2}{\sqrt{\pi}} \operatorname{Erf}\left(\frac{t}{\sqrt{2}}\right) = 1 - \sqrt{\frac{2}{\pi}}\, e^{-\frac{t^2}{2}} \Psi(t)$$

where

$$\Psi(t) \sim \frac{1}{t} - \frac{1}{t^3} + \frac{1\cdot 3}{t^5} - \frac{1\cdot 3\cdot 5}{t^7} + \cdots$$

Now let us return to the problem of tossing the coin. Suppose we agree that an event whose probability is less than 0.01 will be considered to be highly improbable. Then choosing δ so that the right-hand side of (25) should become equal to 0.99 we obtain the corresponding interval which includes all those values which we regard as probable. In our case we obtain $\frac{\delta\sqrt{n}}{\sigma_\xi} = 2.6$. We have $n = 1000$ and $\sigma_\xi = 0.5$, and hence $\delta = 0.041$. Let us denote the number of occurrences of heads by M. Then we get a **confidence interval** $459 \leqslant$ $\leqslant M \leqslant 541$ for M. Thus we have obtained **confidence limits** for the number M which we can guarantee disregarding those events which are considered to be practically impossible relative to the criterion we have chosen (see Sec. 7).

20. Data Processing. In Sec. 19 we considered a random variable ξ whose law of probability distribution was regarded as being known. But in practical problems we most often encounter cases when the law of probability distribution is not known beforehand and when the numerical characteristics of the random variable ξ under investigation should be determined on the basis of the outcomes of the trials. For instance, this is the case when it is necessary to find the mean value of a certain parameter characterizing some property (such as strength or longevity and the like) of a manufactured article belonging to a lot (*general population*) on the basis of inspecting a sample of a number of articles randomly drawn from the lot.

The same problem arises when we repeatedly measure some un-
known physical quantity because the result of every measurement
is a random quantity due to the unavoidable random errors. When
there are no systematic errors the arithmetic mean of the experi-
mental data is approximately taken as the precise value of the
quantity in question and so on.

Suppose that we have performed n trials and that a random variable
ξ has taken the corresponding numerical values $x_1, x_2, \ldots x_n$.
Then our previous investigation indicates that it is the arithmetic
mean value $x^{(n)}$ of the results we have obtained [defined by formu-
la (24)] that should be taken as an approximate value of $\bar{\xi}$. Besides,
it turns out that an approximate value of the standard deviation
σ_ξ can be chosen as

$$\sigma_\xi \approx \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - x^{(n)})^2} \tag{26}$$

To prove (26) we use the notation introduced in Sec. 19 and addi-
tionally put $\zeta_n = \frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \xi^{(n)})^2$. Computing the mean value
of ζ_n we find

$$M\{\zeta_n\} = \frac{1}{n-1} M \left\{ \sum_{i=1}^{n} \xi_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} \xi_i \right)^2 \right\}$$

(check up the result!). Now, representing each ξ_i in the form $\xi_i =$
$= (\xi_i - \bar{\xi}) + \bar{\xi}$ and calculating we obtain

$$\frac{1}{n-1} \left[nD\xi + n\bar{\xi}^2 - \frac{1}{n} (nD\xi + n^2\bar{\xi}^2) \right] = D\xi = \sigma_\xi^2$$

Besides, computing $D\zeta_n$ we find that $D\zeta_n \to 0$ when $n \to \infty$. (Let
the reader perform the calculations.) But the radicand in (26) is
nothing but a value of the random variable ζ_n. This implies (26).

Now we can set a criterion according to which random events will
be considered to be practically impossible (see Sec. 7) and then,
applying formula (25), construct a confidence interval for $\bar{\xi}$. For
instance, let us disregard the events whose probability is less than
0.003. Then (25) indicates that we can put $\frac{\delta\sqrt{n}}{\sigma_\xi} = 3$. Thus, we get
the following confidence interval for $\bar{\xi}$:

$$x^{(n)} - \frac{3\sigma_\xi}{\sqrt{n}} \leqslant \bar{\xi} \leqslant x^{(n)} + \frac{3\sigma_\xi}{\sqrt{n}} \tag{27}$$

The confidence limits thus obtained are guaranteed with a probabi-
lity of 0.997. Formula (27) is widely applied to practical problems.
The value of σ_ξ entering into (27) can be taken from formula (26).

Let us discuss a consequence of formula (27). Suppose that we have performed two independent series of n trials for one and the same random variable ξ and that this has resulted in the mean values $x^{(n)}$ and $\widetilde{x}^{(n)}$. Then estimation (27) holds for both values with the probability $(0.997)^2 = 0.994$, and hence $|x^{(n)} - \widetilde{x}^{(n)}| \leqslant \dfrac{6\sigma_\xi}{\sqrt{n}}$ with this probability. This means that the empirical means $x^{(n)}$ assume some specified values with a great probability in any series of trials when the number of trials n is sufficiently large. This result is quite clear in its qualitative aspect because it is obviously implied by the definition of the mean.

Let us be given two correlated random variables ξ and η and let the regression of both variables be linear (see Sec. 16). If n trials are performed we obtain n pairs of the values assumed by the variables: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Then we can pose the problem of approximating the regression line of η on ξ, on the basis of the data. It turns out that the solution of this problem directly leads to the method of least squares described in Sec. XII.8.

The theory of probability was originated in the 17th century in connection with investigating regularities of games of chance. It was thoroughly developed in the 19th and 20th centuries and is now an important branch of mathematics which has many applications in various divisions of science. Many prominent mathematicians took part in creating the theory of probability. Among Russian scientists who contributed much to the theory we can mention P. L. Chebyshev, A. M. Lyapunov, A. A. Markov, S. N. Bernstein (1880-1968), A. N. Kolmogorov, Yu. V. Linnik and others. There is a mathematical science called **mathematical statistics** which is directly related to the theory of probability. Mathematical statistics deals with problems of processing statistical data. There are many courses on the theory of probability, mathematical statistics and their applications. For a beginner we recommend [1], [17], [41], [45] and [48].

CHAPTER XIX

Computers

The simplest computing devices such as a slide rule, an abacus, an arithmometer and tables are well known and widely applied to practical problems. But these devices do not meet the requirements of modern science, engineering and economics. There are many important problems which are solvable, in principle, and whose solution cannot be practically obtained by means of the above devices because this would take too much time. Therefore in applying these simplest tools we usually disregard many essential factors in order to simplify the calculations, and this often leads to quantitative and even to essential qualitative errors.

The need for more effective computing devices and the achievements of modern technology have led to a revolution in data handling and problem solving practice. This has resulted in inventing and constructing **high-speed electronic computers**. The intensive application of these machines has made it possible to solve many important problems and to obtain fruitful results by means of introducing mathematical methods in a number of new fields of human activities. There is no doubt that the development of modern calculating devices and the extension of their application will result in radical reorganization of scientific research, engineering work, economics, control, service and so on.

§ 1. Two Classes of Computers

There are two basic methods of representing mathematical quantities entering into calculations. The first method is based on a direct representation of mathematical quantities by some physical quantities such as lengths, angles, voltages and the like. To perform certain operations on these quantities it is necessary to construct a physical system in which the corresponding physical quantities are transformed according to the law which describes the transformation of the mathematical quantities in question. Computers based

f which is shown in Fig. 347c (for instance, such a unit can perform the transformation $f(x) \leftarrow \sin x$ if the sine is taken as f and the like), an **amplifier** (Fig. 347d) in which the amplification factor k can be varied etc. In solving differential equations we use a **differentiator** (see Fig. 347e) which produces the output voltage $\frac{dy}{dx}$ for two time-dependent voltages $x(t)$ and $y(t)$ applied to the inputs (in the general case the output $\frac{dy}{dx}$ also depends on t). There are many other units which are used in constructing an analogue computer. Arranging these blocks in different combinations we obtain aggregates capable of solving various equations.

For example, let us consider a scheme intended for solving a system of equations of the form

$$\left. \begin{array}{c} ax + by = c \\ dx + f(y) = 0 \end{array} \right\}$$

It is more convenient to rewrite the system as

$$\left. \begin{array}{c} y = \dfrac{c}{b} - \dfrac{a}{b} x \\ x = -\dfrac{1}{d} f(y) \end{array} \right\}$$

The corresponding diagram is represented in Fig. 348. The circle represents a voltage source with one of its terminals earthed, and
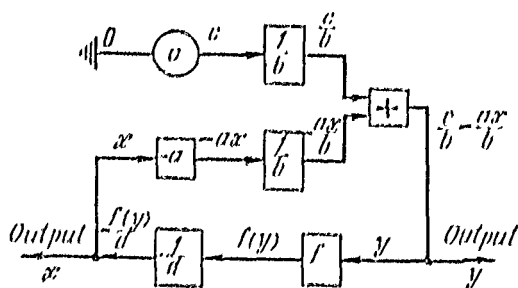
Fig. 348

the voltages transmitted by the channels are also put down there. The scale of the amplification factor can be calibrated in inverse quantities and then there is no need to compute the values of $\frac{1}{b}$ and $\frac{1}{d}$.

We see that this scheme is convenient for studying the way in which the variations of the coefficients affect the solution. Modifying this scheme we can solve many other similar problems. But it should be noted that in case the problem in question has several solutions this scheme may yield a solution different from the one we are interested in.

Let us take one more example illustrating the integration of a differential equation. For instance, let us take an equation of the form

$$y'' + f(y') + y = \psi(x) \quad (y = y(x))$$

The corresponding scheme is shown in Fig. 349. Let the reader verify the correctness of the scheme [in doing this it is more convenient
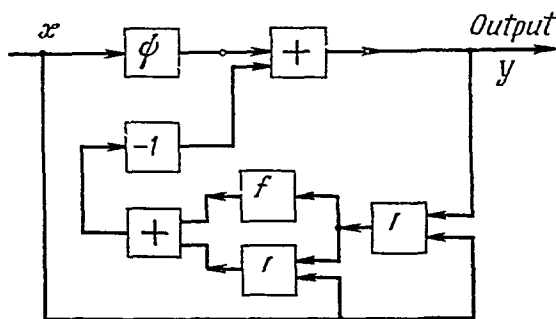
Fig. 349

to rewrite the equation in the form $y = \psi(x) - y'' - f(y')]$. Here the quantity y is a variable whose variations should be related to the variations of the independent variable x.

We often choose the time t as an independent variable. For this case the differentiator (see Fig. 347e) has only one input to which the voltage $\psi(t)$ is applied. The above scheme for solving the differential equation should be then changed correspondingly. Namely, instead of putting in the voltage x we must simply apply the voltage $\psi(t)$ to the terminal which is represented by the point in Fig. 349. This can be performed without calculating $\psi(t)$ if we read this signal from a graph or from some data unit. Such a scheme is particularly convenient when the physical meaning of the problem indicates that the independent variable is the time t. Then the problem can be solved *in real time*. The solution can be directly transmitted (without human intervention) to some other system for a further utilization. Many devices intended for automatic control are based on this idea.

The **real time simulation** makes it possible to replace some expensive aggregates by computers when testing a complicated device. For instance, instead of testing an autopilot in a flight which is

dangerous and expensive we can do it in a test stand where an analogue computer substitutes for a real airplane. It is apparent that this computer must precisely simulate the reaction of the airplane to the performances of the autopilot.

The fundamentals of the theory of analogue computers can be found in [12] and [30].

2. Digital Computers. The first devices of the type of an abacus were invented in China as early as 2,000 or 3,000 years B.C. The first mechanical computer capable of performing addition and subtraction was constructed in 1642 by the prominent French mathematician, physicist and philosopher B. Pascal (1623-1662). The modern desk calculators (summing machines capable of performing addition and subtraction and semiautomatic and automatic arithmometers which perform all the four fundamental operations of arithmetics) are essentially the perfection of Pascal's calculator. All these useful devices cannot work very fast because the speed of performing calculations is low and because the input data are entered into such a machine by an operator.

At the beginning of the 20th century the necessity for processing a large number of similar data in statistics, book-keeping and finances led to the creation of **punch card computers** (tabulators, sorters etc.). The input data to be inserted in such machines are punched on cards (**punch cards**; see Sec. 4). The punch card reader reads the cards by means of a set of brushes (which sense the holes punched in the cards) and produces the corresponding electric pulses. These electric signals make the machine work according to the given program. It can sort the cards, sum up the parameters, accumulate the results and perform some other simple operations. These machines are very useful but they cannot be applied to more complicated calculations.

As it has been mentioned, the revolution in this field is connected with the appearance of a new class of calculating machines which are referred to as **universal high-speed digital automatic computers.** Although the idea of constructing such machines appeared as early as the 19th century, it is only the achievements of modern electronics that have made it possible to realize the idea. The first high-speed electronic digital computer ENIAC (Electronic Numerical Integrator and Computer) was constructed in the USA in 1943. The basic theoretical ideas and principles of constructing such machines were formulated by the American mathematician J. von Neumann (1903-1957) in 1946. In the USSR the first machines of this type were manufactured in 1952.

The *block diagram* of an automatic digital computer is shown in Fig. 350 which represents the main units and their interconnection. In the **memory** (store or storage) there are a number of **locations** (cells or compartments) each of which can store one number or one

instruction (order or command). As we shall see in Sec. 5, the form in which an instruction is represented within the machine does not differ from that of a number. Some of these locations store the information obtained from the input device before the calculations are started whereas the others can be filled in the process of work. When the machine performs calculations the contents of the locations may change many times, and some of the locations may remain empty and may not be used. The control unit interprets the instructions given to the machine and sends the numbers (or instructions) into the arithmetic unit. The arithmetic unit transforms the numbers according to the instructions and then returns them to the memory. As the required results are obtained, they are automatically printed according to the signals sent by the control unit, and after the computations have been completed the control unit stops the computer. (In § 2 we shall describe the sequence of the operations in greater detail.)

Fig. 350

The continuous lines represent information channels and the dotted lines represent control channels

Any mathematical problem can be solved by means of a universal automatic digital computer provided that a certain algorithm is given. An algorithm is understood as an accurate assignment defining a step-by-step procedure for solving the problem which necessarily leads to the required result on the basis of the input data. But there are also many non-mathematical problems for which a similar algorithm can be indicated, and such problems are solvable by means of a digital computer. For example, an electronic computer can control the process of machining a workpiece of a complicated profile in a metal-cutting lathe. In this case the input data characterize the profile, and the results of the calculations are transformed into signals controlling the lathe. The device controlling the flight of an airplane beginning with the take-off and up to the landing at a prescribed place operates in a similar way. A digital computer can control a manufacturing process and makes it possible to completely automatize the process which is particularly important in those industries which are hazardous to human health. If the performance of the process deviates from the initial program the computer can find the most advantageous solution by comparing different variants, and it can also check the result. Of course, the necessary information concerning the real state of the process should be entered in such a machine automatically by means of special devices. A computer can help in designing an engineering construction because it can examine hundreds of variants and choose the best of them by applying a certain given criterion. Digital computers are also

chosen at the end of each location and there are six such digits. Then the maximal number which can be stored in a location (having 30 digits) is represented in the following form:

the sign of the number	the binary digits of the number to the right of the binary point	the sign of the exponent	the binary digits of the exponent

| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

(3)

Hence, this number is equal to $(1 - 2^{-23}) 2^{31} = 2^{31} - 2^8 \approx 2^{31}$. Accordingly, the minimal positive number which can be represented by this method is stored in the form

| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

(4)

and thus it is equal to $2^{-23}2^{-31} = 2^{-54}$ (remember that the first method gave us the least positive number 2^{-29}).

It sometimes occurs that the number of digits in a location is insufficient for guaranteeing the accuracy needed. In such cases it is possible to use a special program which takes two consecutive locations (the first and the second, the third and the fourth etc.) for each number (this is called the *double precision method*).

The input operation, that is "writing" numbers, is realized in different ways depending on the type of computer. The numbers to be introduced into the input unit of a computer (see Sec. 2) are punched on a special tape (**punch tape**) or on a deck of special cards (**punch cards**). In the first case to each location there corresponds a certain part of the tape and in the second case a certain row on a card. The length of a row corresponds to the number of digits contained in a location; a hole punched in the card corresponds to the digit 1 and an unpunched place corresponds to the digit 0. For instance, if the numbers (2), (3) and (4) (we mean the decimal form of the numbers here) follow in succession in the input program the corresponding punched card contains a part of the form shown in Fig. 351 (check it up!). The holes are punched by means of a special perforator (which is not directly connected with the machine) before the machine is started.

A programmer who composes the **program (routine)** for solving a problem often writes the instructions (see Sec. 5) in the octal system and the input numbers in the decimal system. Then after

To represent the subsequent natural numbers we write the corresponding quadruple of the binary digits 0 and 1 in place of each decimal digit. For example, the decimal numbers 63 and 125 are written as:

0110 0011   and   0001 0010 0101

respectively, in this code. Decimal fractions are similarly converted to binary-decimal fractions.

The binary-decimal representation of a number occupies still more space than the binary one but it is very convenient because it is easy to pass to the decimal representation from it and vice versa.

4. **Representing Numbers in a Computer.** In an electronic digital computer numbers are represented in the binary system and stored in the memory unit, storage (see Sec. 2). Each location of the storage in which a number can be stored contains one and the same number of binary places each of which can carry either 0 or 1. There are two main methods of writing numbers in different constructions of computers.

1. *Fixed point method.* In this method the first digit specifies the sign of a number; conventionally, + is represented as 0 and — as 1. The subsequent digits are the binary digits (standing to the right of the binary point) of the binary representation of the number. For example, if there are 30 digits the representation

1	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(2)

corresponds to the binary number —0.001011001 which is expressed as

$$-\left(\frac{1}{8}+\frac{1}{32}+\frac{1}{64}+\frac{1}{128}+\frac{1}{1024}\right)=-\frac{185}{1024}=-0.1806640625$$

in the decimal system.

In such a location (of 30 places) it is possible to write the numbers ranging from $-(1-2^{-29})$ to $+(1-2^{-29})$ with the interval 2^{-29}. Therefore, when putting the necessary quantities into the machine we must supply the numbers which fall out of the range with the corresponding scale factors so that after the multiplication by the factors the numbers should lie in the interval from —1 to +1.

It should be noted that there are some other methods of representing negative numbers but we shall not treat them here.

2. *Floating point method.* This method considerably extends the range of the numbers which can be put into the memory of a machine. It is based on the additional multiplication of a number by a power of two. A certain fixed number of digits is then used for representing the exponent and its sign. Suppose these places are

the program is punched on a keyboard of the perforator (which resembles a cash register used in shops) the instructions are punched on cards in the binary form, according to the program, and the input data are punched in the binary-decimal code. The conversion of numbers to the binary system is carried out according to the instructions written by the programmer in the program, the operation being performed within the machine at the beginning of its work by means of a special **subprogram (subroutine)** furnished into the machine.

Punch cards are in some respects more convenient than a punch tape because a program can be very extensive and can occupy
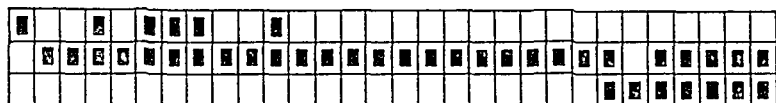
Fig. 351

a deck of cards or a reel of tape, and it is much easier to replace several cards in the deck in case some mistakes have been found than to insert the corrections into the tape.

The representation of numbers stored in the memory of a computer (and transformed in the process of calculations) is realized in the machine by means of **two state elements**. Each location is a set of such elements, the number of the elements being equal to the number of binary digits in the location. A feature of these elements is that each of them can be in one of the two states. One of these states is regarded as a physical embodiment of 0 and the other as a physical embodiment of 1, and thus each binary place in a location can carry either 0 or 1. The construction of the elements varies with the model of a computer but in all cases they must be **inertialess** (i.e. they must pass from one state to the other in a trice) and **stable** (i.e. they must indefinitely long remain in each of the states unless there is a transition signal). An element can be realized as a **trigger** consisting of two triodes whose cathodes are connected together and whose anodes and grids are connected crosswise. In such a circuit either one triode or the other is always cut off, and the transition from one state into the other is performed under the action of a voltage pulse. Charged and uncharged parts of a dielectric plate can also serve as the elements, the transition from one state to the other being performed by means of an electron-beam tube similar to that used in a television set. In some computers magnetic drums are used as storage devices. Such a drum rapidly revolves under a magnetizing head similar to the one used in a tape recorder. The elements in such a device are magnetized and unmagnetized portions on the surface of a drum. **Ferrite cores (magnetic cores)** which can change

the direction of their magnetization when a pulse of current passes through the winding and some other devices are also used as such elements.

A number can be transferred from a location of the storage of a computer to the arithmetic unit and in the opposite direction along a single channel with each digit occurring serially in time sequence, or more than one digit is simultaneously transferred in parallel by means of a system of channels whose number can equal the number of digits in the location. Accordingly, the computers of the first type are referred to as being **serial in storage access** and those of the second type are said to have a **parallel storage system**. The latter method increases the speed of performing operations but at the same time it complicates the construction of a computer.

The greater the number of locations in the storage, the greater the amount of information that can be entered into the machine. Therefore, besides a **high-speed internal memory**, a computer is usually equipped with an **external memory** in order to increase storage capacity. An external memory is usually realized as a magnetic tape from which numbers stored there can be transferred, by means of a special device, to the internal memory. A magnetic tape can carry millions of locations but the speed of recording information on the tape and reading it from the tape is less than that of the internal memory because passing the tape under the magnetizing head requires additional time.

The results of calculations are converted to the decimal system by means of a special subroutine and then printed on paper tape or stored (in the external memory) to be used for further calculations.

Besides handling numerical data a machine can also process some other types of information (for instance, expressed in words) provided it has been coded beforehand by means of binary numbers. It is possible to compile a program which makes a special device decode the results after the operation has been completed, i.e. convert the output data to the form which is natural for this particular type of information. For instance, this is the case in machine translation from one language into another.

5. **Instructions.** A machine performs operations on numbers stored in the memory only according to instructions which are also entered into the memory before the computations are started and which are stored in the binary form not differing from that of numbers. There are **one-address, two-address** and **three-address** instructions which are used depending on the type of the machine. The three-address instructions are especially convenient for programming, and we shall consider them here.

A location storing a three-address instruction can be thought of as being divided into four parts of specified length. For instance,

let these parts contain 3, 9, 9 and 9 binary digits, respectively (as above, we consider locations with 30 digits). The contents of these parts are the following:

(1) The first part termed the **operation part** contains the code of the operation and tells the machine what to do.

(2) The second part carries the **address** (i.e. the number of the location) of the first number taking part in the operation.

(3) The third part contains the address of the second number involved in the operation.

(4) The fourth part stores the address of the location to which the result of the operation must be transferred.

All the locations are numbered in the natural order. If the memory contains 512 locations nine binary digits are sufficient to write the addresses (why?). Suppose that the number 1 is the code of the operation of addition. Then if it is necessary to add together the numbers stored in the 417th and 73rd locations and to transfer the result to the 646th location the corresponding instruction must have the form

$$\underbrace{001}_{} \quad \underbrace{100001111}_{} \quad \underbrace{000111011}_{} \quad \underbrace{110100110}_{}$$
parts: 1st   2nd   3rd   4th   (5)

The reader should verify this form taking into account the above division of locations carrying instructions and the fact that the numbers of locations are coded in the octal system (that is the above numbers 417, 73 and 646 are octal). After the instruction has been executed, the contents of all the locations of the memory (including the 417th and the 73rd locations) remain the same as before except the 646th location in which its former contents are replaced by the sum of the numbers contained in the locations with the addresses 417 and 73.

Instruction (5) is punched on a punch tape or on a punch card and then read in the input unit and transferred to one of the locations in the memory where it is stored in the same way as numbers. The distinction between numbers and instructions only appears in the process of operation of the machine because if the program is compiled correctly the control unit must receive as instructions only those signals which are sent from the locations in which the instructions have been placed.

A one-address instruction is placed in a location divided into two parts which contain the code of an operation and the address of the location. To perform the above operation we need three one-address instructions: the first instruction sends the number from the 417th location into the adder, the second makes the adder sum up the former number and the number stored in the 73rd location and the execution of the last instruction results in transferring the

outcome of the addition to the 646th location. A two-address instruction consists of three parts which carry the code of an operation and the addresses of the numbers on which the operation must be performed. After the operation has been performed the result is sent to the second address.

In what follows we shall consider only three-address instructions. For the sake of simplicity we shall write the signs of operations instead of their codes (for instance, the sign + will designate the code of the operation of addition) and decimal numbers of locations instead of their octal numbers.

In most computers instructions are executed serially, in a consecutive order, unless the program given to the machine contains **transfer instructions** which will be discussed in Sec. 6. Before the machine is started, the program (which is a certain sequence of instructions) and initial data (that is an array of numbers on which the corresponding operations must be performed) are read from punch cards or punch tape and fed into the memory. Then the control unit takes the contents of the first memory location as an instruction and sends the corresponding signals to the arithmetic unit which performs the operations. After that the control unit takes, as the next instruction, the contents of the second location and so on with the exception of transfer instructions mentioned above. When a transfer instruction has been executed the control is transferred not to the subsequent location but to the location whose address is contained in the instruction. Some instructions cause the machine to print the contents of some locations but after the realization of such an instruction the machine also passes to the next location. (By the way, printing takes more time than performing arithmetic operations, and therefore the speed of computations decreases when the program makes the machine print very often.) This step-by-step procedure goes on until the control unit receives the stop instruction after which the machine is brought to a stop. The machine is also automatically stopped if the control unit takes from the program an instruction that cannot be executed (for instance. if a number of which the square root must be extracted turns out to be negative). The same will be in case an *overflow* occurs, that is if the calculations result in a number which is so large that it exceeds the capacity of the location and cannot be written there (see Sec. 4). The control panel is equipped with a special device which makes it possible to check the contents of any location and to insert additional data at any moment of the machine operation process.

Each type of automatic digital computer has its own set of instructions and system of coding. In Sec. 6 we shall give several examples illustrating the main principles of **programming**. The reader should take into account that these examples are given only as an illustration and that in real computers these principles are

realized more economically. For more detail the reader is referred to [8], [18], [20], [21], [25] and [27].

The composition of an extensive program is complicated work which often takes much time and requires much experience.

6. Examples of Programming. Let it be necessary to compute the solution of the system of equations of the first degree

$$\left.\begin{array}{l} ax + by = m \\ cx + dy = n \end{array}\right\}$$

where the numerical values of the quantities a, b, c, d, m and n are considered to be given. According to formulas (VI.2), the sought-for solution is expressed as

$$x = \frac{md - bn}{ad - bc}, \qquad y = \frac{an - mc}{ad - bc} \tag{6}$$

We now place the necessary instructions in locations with addresses (numbers) 1, 2, ... The total number of the required instructions is yet unknown. Let the six input parameters a, b, c, d, m and n be stored in locations having the numbers $\alpha + 1$, $\alpha + 2$, $\alpha + 3$, $\alpha + 4$, $\alpha + 5$ and $\alpha + 6$, respectively. The value of α will be specified later on. Several subsequent locations will be used for storing intermediate results of the calculations. The numbers and instructions stored in these locations before the computations are started do not matter because when a number is written in a location its former contents are automatically erased. The other locations will not be used in compiling the program and calculating the result. Let us begin with computing the first numerator in (6). To compute md we write the instruction

(1) $\times$ $\alpha + 5$ $\alpha + 4$ $\alpha + 7$

(here we shall write the serial number in front of each of the instructions but it should be noted that these numbers are not in fact punched on cards or on tape). After the instruction has been executed the number md is placed in location $(\alpha + 7)$. The next instruction is of the form

(2) $\times$ $\alpha + 2$ $\alpha + 6$ $\alpha + 8$

and its execution results in the appearance of the number bn in location $(\alpha + 8)$. Next, the number bn [stored in location $(\alpha + 8)$] must be subtracted from the number md [location $(\alpha + 7)$]. Since we no longer need these numbers the result can be written in location $(\alpha + 7)$ by means of the following instruction:

(3) — $\alpha + 7$ $\alpha + 8$ $\alpha + 7$

In this example we could have used location $(\alpha + 9)$ for storing the result $md - bn$ but in more complicated programs it is often necessary to economize locations.

Now we similarly put down the instructions which lead to the computation of the denominator of the fractions entering into (6):

(4) $\times$ $\alpha+1$ $\alpha+4$ $\alpha+8$
(5) $\times$ $\alpha+2$ $\alpha+3$ $\alpha+9$
(6) $-$ $\alpha+8$ $\alpha+9$ $\alpha+8$

Further, the numerator stored in location $(\alpha + 7)$ must be divided by the denominator stored in location $(\alpha + 8)$, and the result should be printed:

(7) : $\alpha+7$ $\alpha+8$ $\alpha+7$
(8) Print $\alpha+7$

The execution of the last instruction results in printing the contents of location $(\alpha + 7)$, i.e. the desired value of x, and then the machine proceeds to execute the subsequent instruction. Taking into account that the denominator $ad - bc$ has already been computed and is stored in location $(\alpha + 8)$ we similarly write the instructions which result in computing y and completing the solution of the problem:

(9) $\times$ $\alpha+1$ $\alpha+6$ $\alpha+7$
(10) $\times$ $\alpha+5$ $\alpha+3$ $\alpha+9$
(11) $-$ $\alpha+7$ $\alpha+9$ $\alpha+7$
(12) : $\alpha+7$ $\alpha+8$ $\alpha+7$
(13) Print $\alpha+7$
(14) Stop

The 14th instruction stops the machine. Thus, our program contains 14 instructions and hence we can put $\alpha = 14$. Then the whole program will occupy 20 locations and will have the form

(1) $\times$ 19 18 21
(2) $\times$ 16 20 22
(3) $-$ 21 22 21
(4) $\times$ 15 18 22
(5) $\times$ 16 17 23
(6) $-$ 22 23 22
(7) : 21 22 21
(8) Print 21
(9) $\times$ 15 20 21
(10) $\times$ 19 17 23
(11) $-$ 21 23 21
(12) : 21 22 21
(13) Print 21
(14) Stop
(15) a
(16) b
(17) c
(18) d
(19) m
(20) n

2999 with 2002 and to pass to the execution of the first instruction again, etc. Only after the recurrent addition of unity to the contents of location $(\alpha + 1)$ results in the appearance of 3000 and the inverse of 3050 is computed and printed, the fourth instruction will transfer the control to the next instruction since then we shall have $(\alpha + 3)' = 2999 < (\alpha + 1)' = 3000$. Then the machine must be stopped because all the desired results will have been printed, and therefore the fifth instruction is

(5) Stop

Thus, we can put $\alpha = 5$, and hence the whole program will have the following form:

(1)	$+$	6	7	6
(2)	:	7	6	9
(3)	Print	9		
(4)	Transfer of control	8	6	1
(5)	Stop			
(6)	2000			
(7)	1			
(8)	2999			

We have used only one location 9 for storing the intermediate results but at the same time the contents of location 6 have been changed 1000 times from 2000 to 3000 (with step 1) in the process of the calculations.

We now consider a variant of the above program which results not in printing the reciprocals but in placing them into the locations of the memory with the numbers ranging from 10 to 1009. These stored numbers can be used for some further calculations; of course, we suppose that the memory capacity makes it possible to place the numbers. This program has the following form:

(1)	$+$	6	7	6
(2)	$+$	3	9	3
(3)	:	7	6	9
(4)	Transfer of control	8	6	1
(5)	Stop			
(6)	2000			
(7)	1			
(8)	2999			
(9)	(In this location 1 is the last binary digit, and all the other digits are equal to 0)			

Here we have placed an auxiliary number in location 9 which has no quantitative meaning and serves only for modifying instructions. This is quite a new operation which is performed in the following way in our program: after the second instruction has been executed the first time, the third instruction is sent to the arithmetic unit where it is transformed into

$$: \quad 7 \quad 6 \quad 10$$

and the same part of a program several times. Loops are formed by means of the **conditional transfer instruction** which is of the form

$$\text{Transfer of control} \quad N_1 \quad N_2 \quad N_3$$

(It is apparent that instead of the words "transfer of control" we must in fact·write the code of the operation.) There are many variants of the realization of the instruction. For definiteness, we assume the following interpretation of the above instruction. Let the numbers N_1, N_2 and N_3 designate the addresses of the corresponding locations and let $(N_1)'$, $(N_2)'$ and $(N_3)'$ be the contents of the locations. Suppose that the instruction causes the machine to compare the number $(N_1)'$ with the number $(N_2)'$ and proceed to execute the subsequent instruction if $(N_1)' < (N_2)'$ or the instruction stored in location N_3 if $(N_1)' \geqslant (N_2)'$, the contents of the locations remaining unchanged in either case. In particular, the instruction

$$\text{Transfer of control} \quad 1 \quad 1 \quad N_3$$

causes the machine to execute the instruction stored in location N_3 after the above instruction has been read by the control unit. This is the so-called **unconditional transfer instruction.**

As an example, let us compose a program for printing the table of reciprocals of 1000 successive natural numbers taken, for instance, from 2001 to 3000. Such a program formed without the transfer instruction would be very extensive, but it becomes in fact rather short if the instruction is used. Let us place the number 2000 in location $(\alpha + 1)$ and the number 1 in location $(\alpha + 2)$. Besides. let us place the number 2999 in location $(\alpha + 3)$ (soon we shall see why the number is introduced). Let the first instruction have the form

(1) $+$ $\alpha+1$ $\alpha+2$ $\alpha+1$

(we again place the instructions in locations 1, 2, . . .). After it has been executed the number 2001 substitutes for the number 2000 in location $(\alpha + 1)$. The next two instructions result in calculating the inverse of 2001 and in printing it:

(2) $:$ $\alpha+2$ $\alpha+1$ $\alpha+4$
(3) Print $\alpha+4$

Now we write the instruction

(4) Transfer of control $\alpha+3$ $\alpha+1$ 1

This instruction compares the number stored in location $(\alpha + 1)$ with the number 2999 and transfers the control back to location 1 since the number 2001 placed in location $(\alpha + 1)$ is less than 2999: $(\alpha + 1)' = 2001 < (\alpha + 3)' = 2999$. Thus the machine again executes the first instruction which results in the appearance of the number 2002 instead of 2001 in location $(\alpha + 1)$. Then the instructions (2) and (3) are executed and thus the inverse of 2002 is computed and printed. Next, taking the fourth instruction from the storage and executing it the control unit causes the machine to compare

2999 with 2002 and to pass to the execution of the first instruction
again, etc. Only after the recurrent addition of unity to the contents
of location $(\alpha + 1)$ results in the appearance of 3000 and the inverse
of 3050 is computed and printed, the fourth instruction will trans-
fer the control to the next instruction since then we shall have
$(\alpha + 3)' = 2999 < (\alpha + 1)' = 3000$. Then the machine must be
stopped because all the desired results will have been printed, and
therefore the fifth instruction is

(5) Stop

Thus, we can put $\alpha = 5$, and hence the whole program will have
the following form:

(1)	$+$	6	7	6
(2)	:	7	6	9
(3)	Print	9		
(4)	Transfer of control	8	6	1
(5)	Stop			
(6)	2000			
(7)	1			
(8)	2999			

We have used only one location 9 for storing the intermediate results
but at the same time the contents of location 6 have been changed
1000 times from 2000 to 3000 (with step 1) in the process of the cal-
culations.

We now consider a variant of the above program which results
not in printing the reciprocals but in placing them into the locations
of the memory with the numbers ranging from 10 to 1009. These
stored numbers can be used for some further calculations; of course,
we suppose that the memory capacity makes it possible to place the
numbers. This program has the following form:

(1)	$+$	6	7	6
(2)	$+$	3	9	3
(3)	:	7	6	9
(4)	Transfer of control	8	6	1
(5)	Stop			
(6)	2000			
(7)	1			
(8)	2999			
(9)	(In this location 1 is the last binary digit, and all the other digits are equal to 0)			

Here we have placed an auxiliary number in location 9 which has
no quantitative meaning and serves only for modifying instructions.
This is quite a new operation which is performed in the following
way in our program: after the second instruction has been executed
the first time, the third instruction is sent to the arithmetic unit
where it is transformed into

$$: \quad 7 \quad 6 \quad 10$$

[By the way, it should be noted that the sign $+$ in instructions (1) and (2) designates the operations of addition which are performed according to different rules and therefore they have different codes. For more detail the reader is referred to the books enumerated above.] The third instruction taking the above form, its execution results in placing the number $\frac{1}{2001}$ in location 10. After the repeated execution of the second instruction, the third instruction is modified again and takes the form

$$: \quad 7 \quad 6 \quad 11$$

Therefore the second execution of instruction (3) causes the control unit to send the number $\frac{1}{2002}$ to location 11 and so on. Thus we have encountered here the operation of modifying an address entering into an instruction.

Hence, it is possible to perform operations on instructions thus automatically modifying them in the process of work of the machine. This obviously extends the application of digital computers.

The conditional transfer instruction is used not only in employing loops but also when it is necessary to introduce the **branching instruction** which causes the computer to perform different sequences of operations depending on some circumstances unknown beforehand. For example, suppose that a number a must appear in a certain location β, and let it be necessary to retain it unchanged if $a \geqslant 0$ and square it and store the result in the same location β if $a < 0$. In the zeroth location the number 0 is usually placed, and therefore we can realize the desired procedure by writing in the corresponding place of the program the following instructions:

$$\cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot \ \cdot$$

(k)	Transfer of control	β	0	$k+2$
$(k+1)$	$\times$	β	β	β

Branching is then performed automatically, and when the program has been executed we shall not even know which variant has been realized unless a special instruction causing the machine to output the information concerning this procedure is introduced into the program.

Finally, let us consider an example of a program in which the total number of operations is not set beforehand. Let it be necessary to solve the cubic equation

$$[x = 0.1x^3 + 1]$$

by applying the iterative method (see Sec. V.3) and beginning with the initial approximation $x_0 = 0$. To do this we place the number 0 in location $(\alpha + 1)$ (this practically means that the corresponding

location $(\alpha + 2)$ and the number 1 in location $(\alpha + 3)$. The successive approximations appearing in the process of calculations will be placed in location $(\alpha + 1)$.

After a certain approximation has been computed the calculations yielding the next approximation can be carried out according to the following instructions:

(1)	$\times$	$\alpha + 1$	$\alpha + 1$	$\alpha + 4$
(2)	$\times$	$\alpha + 4$	$\alpha + 1$	$\alpha + 4$
(3)	$\times$	$\alpha + 2$	$\alpha + 4$	$\alpha + 4$
(4)	$\div$	$\alpha + 4$	$\alpha + 3$	$\alpha + 4$

Thus. the new approximation will be placed in location $(\alpha + 4)$. It must be compared with the preceding approximation stored in location $(\alpha + 1)$. If the approximations differ the result must be transferred to location $(\alpha + 1)$ and then the iteration should be repeated. If the approximations coincide the result must be printed and the machine must be stopped. This can be realized by means of the following instructions (check!):

(5) $|-|$ $\alpha + 1$ $\alpha + 4$ $\alpha + 1$

[the execution of this instruction results in sending the absolute value of the difference between the contents of locations $(\alpha + 1)$ and $(\alpha + 4)$ to location $(\alpha + 1)$]

(6) Transfer of control	0	$\alpha + 1$	9
(7)	$\div$	$\alpha + 4$	0	$\alpha + 1$
(8) Transfer of control	1	1	1
(9) Print	;	$\alpha + 4$
(10) Stop

Consequently, we can put $\alpha = 10$ and write down the whole program which occupies 13 locations. The program will cause the machine to perform the iterations until a subsequent approximation coincides with the preceding one (note that if the iterative process converges this aim is necessarily achieved because the results of the calculations are automatically rounded off). Then the machine will print the result and stop. It should be noted that if the iterative process does not converge either an overflow of locations will occur or the machine will go into a closed loop. that is repeat a certain sequence of instructions over and over again. In the latter case the machine cannot stop on its own, and we must stop it by pressing the stop key on the keyboard.

Programs for solving more complicated problems can be very extensive. But they often include some simpler problems, for instance, such as the problem of computing the value of the sine for a given value of the argument entering into the main problem and so on. These simpler problems are encountered very often and it is therefore expedient to have special subroutines for solving them. These subroutines are composed beforehand and stored in certain locations in the external or internal memory. When a problem of

this kind is encountered in solving a more complicated problem a special instruction is introduced into the program which makes the control unit take the corresponding subroutine out of the memory.

To facilitate programming, several programming languages for communicating with computers have been recently developed. A program can be written in these languages so that all the necessary procedures can be readily and accurately expressed in terms which are closer to the language of mathematics than those described above. A program written in such a language is independent of the type of the machine we use and is printed by means of a device resembling an ordinary typewriter. Then the program is automatically translated into the machine code by means of a special processor. Among these languages we mention the FORTRAN language (FORmula TRANslating system) and the ALGOL 60 (ALGOrithmic Language developed in 1960). The introduction of the languages in practice will make the application of computers available for many scientists and engineers. By the way, in some cases an experienced programmer can compose a program in the machine code more economically without using a computer language because he takes into account some peculiarities of the concrete machine. This is especially important when we have to economize machine time.

The errors occurring in a calculation process can appear because of the mistakes in composing or punching the program or due to malfunctions of the machine itself. The latter can be systematic (for instance, when some elements get out of order) or random (when an element passes from one state to the other on its own, at random). Systematic errors are checked by means of built-in checks or supplementary programmed checks based on solving certain problems with the answers known in advance. The correctness of composing a program which is to be introduced into the machine is checked before the machine is started. When checking a program we usually make the machine execute some parts of the program and try to roughly estimate certain intermediate results which is very important because these results must not exceed the capacity of locations of the memory (see Sec. 4). To check the punching we usually repeatedly punch the program and cause the machine to compare automatically the corresponding cards from the two decks with one another. To detect random malfunctions we can apply some well-known arithmetical rules used for checking the results of calculations. In more important cases the result is computed repeatedly in order to compare the answers. Besides, computer designers try to eliminate the possibility of random malfunctions by perfecting the machine in the process of designing, manufacturing and modifying its components.

Equations of Mathematical Physics

Equations of mathematical physics are mainly partial differential equations describing various physical processes. The theory of these equations is an important division of mathematics with many applications to physics, engineering and other branches of science. Here we shall present some elementary facts of the theory.

§ 1. Classical Equations of Mathematical Physics

1. Derivation of Some Equations. Let us consider the process of *longitudinal vibrations of an elastic rectilinear homogeneous bar.* We shall draw the x-axis along the bar and denote by x the coordinate of the corresponding point of the bar in the state of equilibrium when the bar is unloaded. Let $u = u\,(x,\ t)$ be the longitudinal displacement of the point x at moment t. In investigating the phenomenon we shall assume the **hypothesis of plane sections**, that is we shall suppose that the plane sections of the bar move in the process of vibration in such a way that they remain parallel to their initial positions all the time (practically this means that the deviations from this condition are inessential and can be disregarded).

The function $u\,(x,\ t)$ determines the law of vibrations of the bar. It should be noted that the term "vibrations" is understood in a conditional sense here because the process of deformation of the bar is an oscillatory motion only in certain simpler cases whereas in the general case the motion can be of a more complicated nature.

In a vibration process there appear elastic stresses in the cross sections of the bar. We shall suppose that the corresponding deformations lie within the limits of applicability of well-known **Hooke's law** (discovered in 1660 by the English mathematician and inventor R. Hooke, 1635-1703). The law states that unless a certain limit is exceeded the normal stress σ is directly proportional to the longitudinal elongation ε, i.e. $\sigma = E\varepsilon$ where E is **Young's modulus** characterizing the properties of the material. Therefore the stress can

be readily expressed in terms of the function u because the longitudinal elongation of an element dx of the bar is equal to $\frac{\partial_x u}{\partial x} = \frac{\partial u}{\partial x}$ (see Fig. 352) and thus we have

$$\sigma = E \frac{\partial u}{\partial x} \qquad (1)$$

Now we can write down the equation of motion of the element dx of the bar. Let us suppose, for generality, that the bar is subjected
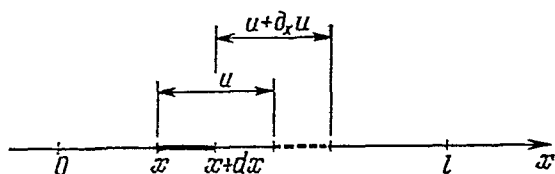
Fig. 352
—— element dx in free equilibrium state
— — — the same element in the process of vibrations

to a longitudinal distributed load of intensity $p = p(x, t)$. Let F denote the constant cross-section area of the bar. Then the resultant force acting upon the element is equal to

$$[\sigma F + \partial_x (\sigma F)] - \sigma F + p\,dx = F \frac{\partial \sigma}{\partial x} dx + p\,dx = \left(FE \frac{\partial^2 u}{\partial x^2} + p \right) dx$$

Denoting the density of the material of the bar by ρ we write, on the basis of Newton's second law, the equation

$$\left(FE \frac{\partial^2 u}{\partial x^2} + p \right) dx = \rho F\,dx \frac{\partial^2 u}{\partial t^2}$$

Finally, dividing by $\rho F\,dx$ and introducing the notation

$$\sqrt{\frac{E}{\rho}} = a, \qquad \frac{p(x, t)}{\rho F} = f(x, t)$$

we derive the equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} + f(x, t) \qquad (2)$$

where $f(x, t)$ is the given function and $u = u(x, t)$ is the sought-for one.

Let us derive another important equation, namely, the heat equation describing the process of propagation of heat in a medium. We shall restrict ourselves to the case of a homogeneous isotropic medium. Let us denote by $u = u(x, y, z, t)$ the temperature at the point (x, y, z) of the medium at moment t. In deducing the equation which must be satisfied by the function u we shall apply the Coulomb law which states that at each point of a medium the thermal

energy is transferred in the direction of $-\operatorname{grad} u$ (which means that the heat propagates from the regions of higher temperature to those of lower temperature) and that the speed of the propagation is proportional to $|\operatorname{grad} u|$. This law holds with a sufficient accuracy when the variations of the temperature are comparatively small. To put down the mathematical expression of the law we take into account that it means that the quantity of heat passing through a surface element (dS) during the time period dt is equal to

$$dQ = k \operatorname{grad}_\mathrm{n} u \, dS \, dt$$

where k is the **coefficient of heat conduction** and n is the outer unit normal vector to the area. Then the total quantity of heat passing during time dt into the interior of a solid (Ω) with boundary surface (S) is equal to

$$k \oint_{(S)} \operatorname{grad}_\mathrm{n} u \, dS \, dt = k \oint_{(S)} \operatorname{grad} u \cdot d\mathbf{S} \, dt$$

· (compare this with Sec. XVI.22). Applying Ostrogradsky's formula (see Sec. XVI.23) and considering the volume (Ω) to be small we obtain (compare with Sec. XVII.22), to within infinitesimals of higher order, the relation

$$dQ = k \int_{(\Omega)} \operatorname{div} \operatorname{grad} u \, d\Omega \, dt =$$

$$= k \int_{(\Omega)} \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) d\Omega \, dt = k \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) \Omega \, dt$$

Further we shall write $(d\Omega)$ instead of (Ω).

Now we proceed to write the equation of heat balance for the element of volume $(d\Omega)$. Let us suppose, for generality, that there are sources of thermal energy in the medium distributed with density $q = q(x, y, z, t)$. The quantity of heat in the volume $(d\Omega)$ is equal to $c\rho u \, d\Omega$ where c is the **specific heat** and ρ is the density of the substance. It follows that

$$\partial_t (c\rho u \, d\Omega) = k \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) d\Omega \, dt + q \, d\Omega \, dt$$

Cancelling out $c\rho \, d\Omega \, dt$ in both sides and introducing the notation $\frac{k}{c\rho} = a$ (a is the **coefficient of temperature conductivity**) and $\frac{q}{c\rho} = f$ we arrive at the **heat equation**

$$\frac{\partial u}{\partial t} = a \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) + f(x, y, z, t) \tag{3}$$

where the function $f(x, y, z, t)$ is given and $u = u(x, y, z, t)$ is the sought-for function.

If there are no sources of thermal energy in the part of space under consideration equation (3) takes the simplified form

$$\frac{\partial u}{\partial t} = a \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) \qquad (4)$$

2. Some Other Equations. It turns out that various physical processes are described by means of similar differential equations.

For instance, equation (2) describes not only the process of longitudinal vibrations of a bar but also the process of small transverse oscillations of a taut string (in this case u denotes the transverse deflection; see Sec. XVII.31), the process of longitudinal oscillations of a gas in a tube (in this case u is the pressure or the density), the oscillations of an electric current in a wire with distributed resistance and inductance when there are no energy losses etc.

The equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \qquad (5)$$

is of particular importance. It describes **free** one-dimensional oscillations of a homogeneous medium, whereas equation (2) describes **forced** oscillations. In the case of a non-homogeneous medium the equation becomes more complicated.

These equations should be accordingly changed in the case of two-dimensional and three-dimensional vibration processes. For instance, the equation of vibrations of a homogeneous isotropic three-dimensional medium is of the form

$$\frac{\partial^2 u}{\partial t^2} = a^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) + f(x, y, z, t) \qquad (6)$$

in the case of forced oscillations and of the form

$$\frac{\partial^2 u}{\partial t^2} = a^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right)$$

in the case of free oscillations. These equations are also satisfied by the projections of the vectors of electric and magnetic field intensities, by the projections of the displacement vector of elastic vibrations of a body and so on.

Equations (3) and (4) are of another type. We have seen that these equations are satisfied by the temperature in the process of heat transfer in an isotropic homogeneous medium. It can be shown that the same equations describe the diffusion processes if u denotes the density of a diffusing substance. Equations (3) and (4) can also be considered in a plane or on a straight line. For instance, in the latter case the equations take, respectively, the forms

$$\frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2} + f(x, t) \qquad (7)$$

and

$$\frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2} \tag{8}$$

If the process in question is **stationary** both the sought-for function u and the function f describing an external action must be independent of t. Then from (6) we obtain the equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = -\frac{1}{a^2} f(x, y, z) = f_1(x, y, z) \tag{9}$$

which is referred to as **Poisson's equation**. When, in a part of space, there are no external actions, we obtain the equation

$$\frac{\partial^2 u}{[\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0 \tag{10}$$

called **Laplace's equation**. Equations (3) and (4) yield the same equations (9) and (10) in the stationary case.

Here we have put down only those equations of mathematical physics which are most thoroughly studied. There are many other equations describing various phenomena.

3. Initial and Boundary Conditions. We showed in Sec. XV.2 that every ordinary differential equation possesses infinitely many solutions and that to isolate a concrete solution we need some subsidiary conditions. The same is true for partial differential equations for which we usually set initial or boundary conditions or both as the subsidiary conditions.

We now return to the problem of longitudinal vibrations of a bar (see Sec. 1). It seems natural that to obtain a concrete process of vibrations we must set the initial conditions specifying the initial state of the bar. From the point of view of physics the initial state is completely determined by the corresponding initial displacements and initial velocities of the points of the bar [compare with conditions (XV.10)]. Thus, for equation (2), we write the **initial conditions**

$$u\,|_{t=0} = \varphi(x) \quad \text{and} \quad \frac{\partial u}{\partial t}\bigg|_{t=0} = \psi(x) \tag{11}$$

where the functions φ and ψ are given [compare with formulas (XVII.133)].

If we investigate a portion of the bar which is placed so far from its ends that their influence is inessential during the time period in question (in other words, if it is possible to regard the bar as being infinite) it is sufficient to know only initial conditions (11) in order to determine the solution. In this case we arrive at a problem with initial conditions alone, i.e. **Cauchy's problem**. The initial conditions and Cauchy's problem are of a similar type for equation (6) in a plane or in space [equation (6) is referred to as the **wave equation**]. The initial conditions for heat equation (3) or (7) differ

from those of (6) because the physical meaning of the problem indicates that to specify uniquely a process of heat transfer it is sufficient to specify only the initial distribution of temperature and therefore it is only the initial values of u that should be set for $t = 0$.

If the influence of the ends of the bar cannot be neglected then besides the initial conditions we must set certain **boundary conditions** which describe the processes on the boundary of the medium in question, i.e. at the ends of the bar in our case. Boundary conditions can be of different forms depending on the processes, and they should be set for both ends ($x = 0$ and $x = l$) independently. For instance, let us consider the left end.

It can be rigidly fixed and then we have

$$u \mid_{x=0} = 0 \quad (\text{i.e. } u(0, t) = 0) \quad . \tag{12}$$

A more general condition can describe the motion of the left end according to a given law of the form

$$u \mid_{x=0} = \chi(t) \tag{13}$$

where the function $\chi(t)$ is given. Condition (13), and its special case (12), is called a *condition of the first kind*.

If the left end is free the condition must express the fact that there is no normal stress there. Thus, by (1), we have

$$\frac{\partial u}{\partial x} \bigg|_{x=0} = 0$$

This condition [and also a more general condition $\frac{\partial u}{\partial x}\big|_{x=0} = \chi(t)$] is called a *condition of the second kind*.

In the case of an **elastic fixing** of the left end the normal stress in the end point section is proportional to its displacement, that is we have $\sigma = ku$ at the end-point. From this, on the basis of formula (1), we deduce the condition

$$\left(\frac{\partial u}{\partial x} - \frac{k}{E} u \right) \bigg|_{x=0} = \left(\frac{\partial u}{\partial x} - \alpha u \right) \bigg|_{x=0} = 0 \quad \left(\alpha = \frac{k}{E} \right) \tag{14}$$

which is referred to as a *condition of the third kind*; the same term is applied to the corresponding non-homogeneous condition. (Let the reader verify that in the case of the right end the similar condition contains $+$ instead of $-$ in front of α.)

Thus, if, for instance, the left end of the bar is rigidly fixed and the right end is free the boundary conditions are of the form

$$u \mid_{x=0} = 0, \quad \frac{\partial u}{\partial x} \bigg|_{x=l} = 0$$

The problem of solving a partial differential equation when both initial and boundary conditions are given is referred to as a **mixed problem (boundary-initial-value problem)**.

If an equation is solved in a three- or two-dimensional domain a boundary condition of the first kind reduces to specifying the values of the sought-for function on the boundary of the domain and a condition of the second kind consists in setting the values of the derivative of the function along the normal to the boundary.

Thus, if the influence of the boundary of the domain is essential the equation of the **non-stationary process** in question must be solved for given initial and boundary conditions. It is apparent that in the case of the equation of a **stationary process** [for instance, equation (9) or (10)] we set only boundary conditions. In the latter case, for a condition of the first kind, the problem is called the **Dirichlet problem** or the **first boundary value problem** (after the German mathematician P. Dirichlet, 1805-1859). Accordingly, for a condition of the second kind it is called the **Neumann problem** or the **second boundary value problem** after the German mathematician K. Neumann (1832-1925) who for the first time systematically investigated the problem in 1877.

§ 2. Method of Separation of Variables

There are many methods of solving equations of mathematical physics. Here we shall discuss one of the most important methods referred to as the **method of separation of variables**.

4. Basic Example. Let us consider the problem of solving the equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \quad (0 \leqslant x \leqslant l, \quad 0 \leqslant t < \infty) \tag{15}$$

for the simplest boundary conditions

$$u\,|_{x=0} = 0, \quad u\,|_{x=l} = 0 \quad (0 \leqslant t < \infty) \tag{16}$$

and initial conditions

$$u\,|_{t=0} = \varphi(x), \quad \frac{\partial u}{\partial t}\bigg|_{t=0} = \psi(x) \quad (0 \leqslant x \leqslant l) \tag{17}$$

where φ and ψ are some given functions.*

We shall interpret the function $u = u(x, t)$ as the transverse deflection of a taut vibrating unloaded string (see Sec. XVII.31) fixed at the ends ($x = 0$ and $x = l$) and the functions φ and ψ in conditions (17) as an initial deflection and an initial velocity. We can, of course, interpret $u(x, t)$ as a longitudinal displacement of the point x of a bar and the like.

---

* In Sec. XVII.31 we solved this problem [see equation (121) and conditions (132) and (133)] by means of expansions in Fourier series. The method of separation of variables, as we shall see, reduces to Fourier expansions and yields the same result [see formulas (25) and (XVII.135)].—*Tr.*

The main idea of the method of separation of variables is that we seek for a solution of equation (15) of the special form

$$u (x,\ t) = X (x)\ T (t) \qquad (18)$$

under boundary conditions (16) (but without any initial conditions). Hence, we are interested in a solution which is the product of a function dependent only on x by a function dependent only on t. If we want to investigate the form of the string corresponding to solution (18) at the subsequent moments t_1, t_2, t_3, . . . we must multiply the fixed function $X (x)$ by the constant factors $T (t_1)$, $T (t_2)$, $T (t_3)$, ... Therefore the zeros of the function $X (x)$ remain all the time the zeros of the function $u (x, t)$ (and are referred to as the **nodes**). Accordingly, the points of extrema of the function $X (x)$ remain the points of extrema of $u (x, t)$

Fig. 353

(these are called the **antinodes**). The form of the vibrating string at the successive moments of time is depicted in Fig. 353. A vibrational state of peculiar type (18) is called a **standing wave**. The function $X (x)$ describes the form of the standing wave and $T (t)$ expresses the law of its variation in time. Hence, we have posed the problem of finding the standing waves which are possible for the given boundary conditions.
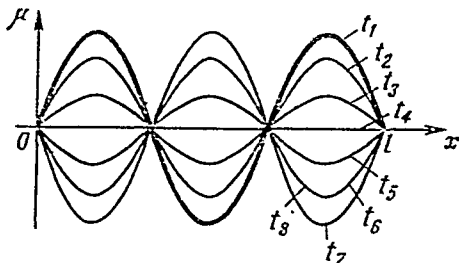
The substitution of (18) into (15) results in

$$X (x)\ T'' (t) = a^2 X'' (x)\ T (t), \quad \text{i.e.} \quad \frac{T'' (t)}{a^2 T (t)} = \frac{X'' (x)}{X (x)}$$

The left-hand side being independent of x and the right-hand side being independent of t, the last equality can be fulfilled if and only if both sides contain neither x nor t. Hence, both sides are equal to the same constant which we denote by $-\lambda$ (but λ itself can be of an arbitrary sign):

$$\frac{T'' (t)}{a^2 T (t)} = \frac{X'' (x)}{X (x)} = -\lambda$$

It follows that $T (t)$ and $X (x)$ must satisfy the corresponding ordinary differential equations

$$T'' (t) + \lambda a^2 T (t) = 0 \qquad (19)$$

and

$$X'' (x) + \lambda X (x) = 0 \qquad (20)$$

Thus, the independent variables have been separated!

50*

Now recall that we have imposed conditions (16) on solution (18). The first condition (16) yields $X(0) \, T(t) \equiv 0$ which implies $X(0) = 0$ [if $X(0) \neq 0$ we must have $T(t) \equiv 0$ and then $u(x, t) \equiv 0$ which does not yield a standing wave]. We similarly consider the second condition (16) and thus arrive at the boundary conditions for a standing wave:

$$X(0) = 0, \quad X(l) = 0 \qquad (21)$$

Consequently, to find all the possible forms of standing waves we must solve equation (20) with boundary conditions (21) (a similar problem was treated in Sec. XV.16 but here we shall not use the results obtained there). It is clear that the function $X \equiv 0$ satisfies both equation (20) and conditions (21) for any λ but we are not interested in the function since it does not yield a standing wave. Thus, the only solutions we are interested in are those satisfying the condition $X(x) \not\equiv 0$. Such solutions, generally speaking, may not exist for all λ. The values of λ for which these solutions exist are called the **eigenvalues** of problem (20) with conditions (21). The solutions $X(x)$ are called the **eigenfunctions** of the problem corresponding to these eigenvalues.

We first suppose that $\lambda < 0$, i.e. $\lambda = -\nu^2$. Then equation (20) has the general solution

$$X(x) = C_1 e^{\nu x} + C_2 e^{-\nu x}$$

(check it up!), and conditions (21) imply that

$$C_1 + C_2 = 0 \quad \text{and} \quad C_1 e^{\nu l} + C_2 e^{-\nu l} = 0$$

It follows that $C_2 = -C_1$ and therefore

$$C_1 e^{\nu l} - C_1 e^{-\nu l} = 0, \quad \text{i.e.} \quad C_1 (e^{2\nu l} - 1) \, e^{-\nu l} = 0$$

But the second and the third factors are different from zero (why?). Hence $C_1 = 0$ and consequently $C_2 = 0$ and $X(x) \equiv 0$. Thus, there are no negative eigenvalues of this problem. Let the reader prove that the value $\lambda = 0$ is not an eigenvalue either.

Now let $\lambda = k^2 > 0$. Then equation (20) has the general solution

$$X(x) = C_1 \cos kx + C_2 \sin kx \qquad (22)$$

(check it up!). Conditions (21) imply that

$$C_1 = 0 \quad \text{and} \quad C_2 \sin kl = 0 \qquad (23)$$

But we must have $C_2 \neq 0$ (why?) and hence $\sin kl = 0$. From this we deduce

$$kl = n\pi \quad \text{and} \quad k = \frac{n\pi}{l} \quad (n = 1, 2, 3, \ldots)$$

Thus, the problem has the following infinite set (spectrum) of eigenvalues:

$$\lambda = \lambda_n = \left(\frac{n\pi}{l}\right)^2 \quad (n = 1,\ 2,\ 3,\ \ldots)$$

The corresponding eigenfunctions are implied by (22):

$$X_n(x) = \sin \frac{n\pi x}{l} \quad (n = 1,\ 2,\ \ldots) \tag{24}$$

We have not put down the factor C_2 here because any constant factor can be included into $T(t)$ in the expression (18).

Substituting the value $\lambda = \lambda_n$ $(n = 1,\ 2,\ 3,\ \ldots)$ thus found into equation (18) we get

$$T(t) = A \cos \frac{an\pi}{l} t + B \sin \frac{an\pi}{l} t$$

where A and B are arbitrary constants. Thus, we have obtained harmonic vibrations with frequency $\omega_n = \frac{an\pi}{l}$. Consequently,

Fig. 354

according to (18), the sought-for standing waves are of the form

$$u(x,\ t) = \left(A \cos \frac{an\pi}{l} t + B \sin \frac{an\pi}{l} t\right) \sin \frac{n\pi x}{l} \quad (n = 1,\ 2,\ \ldots)$$

(The first three standing waves corresponding to $n = 1, 2, 3$ are shown in Fig. 354.) The frequencies of vibrations of these waves are equal to

$$\omega_1 = \frac{a\pi}{l}, \quad \omega_2 = \frac{2a\pi}{l} = 2\omega_1, \quad \omega_3 = \frac{3a\pi}{l} = 3\omega_1, \quad \ldots$$

As it is said in acoustics, the first standing wave corresponds to the **fundamental tone** and the subsequent standing waves whose frequencies are 2, 3, 4 etc. times the frequency of the fundamental tone determine the **overtones** (see Sec. XVII.23).

Let us now construct the general solution of equation (15) with boundary conditions (16) which must describe the general form of vibrations that are possible under the given boundary conditions. This general solution is obtained as a combination (*superposition*) of the above standing waves with different amplitudes, i.e. it has the form

$$u(x,\ t) = \sum_{n=1}^{\infty} \left(A_n \cos \frac{an\pi}{l} t + B_n \sin \frac{an\pi}{l} t\right) \sin \frac{n\pi x}{l} \tag{25}$$
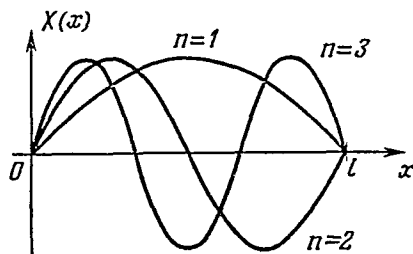
where A_n and B_n ($n = 1, 2, 3, \ldots$) are arbitrary constants. Equation (15) and boundary conditions (16) being linear and homogeneous, the whole sum (25) satisfies them because each summand satisfies the equation and the conditions (compare with property 1 in Sec. XV.14). To prove that formula (25) represents the general solution we must show that the arbitrary constants A_n and B_n can be so chosen that any initial conditions (17) should be satisfied. Note that in contrast to an ordinary differential equation whose general solution contains a finite number of arbitrary constants the general solution of problem (15), (16) depends on two *infinite* sequences of arbitrary constants or, which is the same, on two arbitrary functions, namely, the functions $\varphi(x)$ and $\psi(x)$ entering into conditions (17).

To satisfy the first condition (17) we put $t = 0$ in (25) which yields

$$\varphi(x) = \sum_{n=1}^{\infty} A_n \sin \frac{n\pi x}{l} \quad (0 \leqslant x \leqslant l) \tag{26}$$

Thus, the given function $\varphi(x)$ must be expanded into a series in eigenfunctions of problem (20), (21). In this particular case this is nothing but an expansion into a Fourier series of form (XVII.107). As it was shown in Sec. XVII.22, expansion (26) is possible, and the coefficients of the expansion are found on the basis of the orthogonality condition:

$$A_n = \frac{\int_0^l \varphi(x) \sin \frac{n\pi x}{l} \, dx}{\int_0^l \sin^2 \frac{n\pi x}{l} \, dx} = \frac{2}{l} \int_0^l \varphi(x) \sin \frac{n\pi x}{l} \, dx \tag{27}$$

To satisfy the second condition (17) we differentiate both sides of equality (25) with respect to t and then put $t = 0$:

$$\psi(x) = \sum_{n=1}^{\infty} B_n \frac{an\pi}{l} \sin \frac{n\pi x}{l}$$

After a manner of (27) we obtain the relation

$$B_n = \frac{2}{an\pi} \int_0^l \psi(x) \sin \frac{n\pi x}{l} \, dx \tag{28}$$

(check up the calculations!).

Thus, the solution of the original problem (15)-(17) is given by formula (25) in which the coefficients are defined by formulas (27) and (28).

5. Some Other Problems.

1. Suppose that instead of equation (15) we consider homogeneous heat equation (8) with the same simplest boundary conditions (16) and with initial conditions of the form

$$u \mid _{t=0} = \varphi (x) \quad (0 \leqslant x \leqslant l)$$

(see Sec. 3). After substitution (18) has been performed and the variables have been separated we arrive at the same problem (20), (21) whose solution yields the eigenfunctions and the eigenvalues (let the reader perform the calculations!). But instead of (19) we obtain the equation

$$T' (t) + \lambda a T (t) = 0$$

for the function $T (t)$. It follows that

$$T (t) = A e^{-\lambda a t}$$

Therefore we obtain the formula

$$u (x, t) = \sum_{n=1}^{\infty} A_n e^{-\frac{an^2\pi^2}{l^2} t} \sin \frac{n\pi x}{l}$$

which expresses the general solution of the equation under the given boundary conditions. This formula substitutes for expression (25) in this case. The coefficients A_n can be found here by means of the same formula (27).

We see that the set of the eigenfunctions and eigenvalues remains the same as before in this problem but the dependence on time is of a different type here. Instead of harmonic functions of , which we had in Sec. 4 we obtain exponentially damping functions in this case (which results in $\lim\limits_{t \to \infty} u = 0$). By the way, the physical meaning of the problem in question implies that the function $u (x, t)$ must behave in this peculiar way as t increases (why?).

Now take the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad (0 \leqslant x \leqslant l, \quad 0 \leqslant y \leqslant m)$$

with the conditions

$$u \mid _{x=0} = 0, \quad u \mid _{x=l} = 0 \quad (0 \leqslant y \leqslant m)$$

Let the reader verify that the same method yields the general solution of the form

$$u (x, y) = \sum_{n=1}^{\infty} \left(A_n e^{\frac{n\pi y}{l}} + B_n e^{-\frac{n\pi y}{l}} \right) \sin \frac{n\pi x}{l}$$

for this problem. The constants A_n and B_n can be specified if some subsidiary boundary conditions for $y = 0$ and $y = m$ are given (then we substitute $y = 0$ and $y = m$ into the above solution etc.).

It should be noted that such a separation of variables is by far not always possible for all the partial differential equations.

2. We now turn back to equation (15). Suppose that instead of boundary conditions (16) we have conditions of another type. For instance, let us take the conditions

$$u\,|_{x=0} = 0, \quad \frac{\partial u}{\partial x}\Big|_{x=l} = 0$$

This results in a change of conditions (21) specifying the eigenfunctions. The new conditions will be of the form

$$X(0) = 0, \quad X'(l) = 0$$

Then instead of (23) we arrive at the relations

$$C_1 = 0, \quad C_2 k \cos kl = 0$$

which imply

$$kl = -\frac{\pi}{2} + n\pi \quad (n = 1, 2, \ldots)$$

Therefore the eigenvalues and the eigenfunctions of this problem are expressed by the formulas

$$\lambda_n = \frac{1}{l^2}\left(-\frac{\pi}{2} + n\pi\right)^2, \quad X_n(x) = \sin\left[\left(-\frac{\pi}{2} + n\pi\right)\frac{x}{l}\right]$$
$$(n = 1, 2, 3, \ldots)$$

Hence, the spectrum of eigenvalues (together with the spectrum of frequencies) and the set of eigenfunctions have changed. The first four eigenfunctions are depicted in Fig. 355. It is easy to directly verify that the eigenfunctions are orthogonal to one another. It can be proved that the eigenfunctions of the problems of this kind always form a complete system of functions. Therefore the solution of such problems can be completed by means of techniques similar to those given in Sec. 4.

We often deal with more complicated equations when finding eigenfunctions of a problem. For instance, take the boundary conditions of the form

$$u\,|_{x=0} = 0, \quad \left(\frac{\partial u}{\partial x} + \alpha u\right)\Big|_{x=l} = 0$$

[see formula (14)]. Then instead of (23) we obtain

$$C_1 = 0, \quad C_2(k \cos kl + \alpha \sin kl) = 0$$

which implies $\tan kl = -\dfrac{k}{\alpha}$. Let us denote kl by μ. The equation for determining μ is of the form $\tan \mu = -\dfrac{\mu}{\alpha l}$. A method of graphical solution of the equation is illustrated in Fig. 356. The graphs

clearly indicate that there is an infinite sequence of positive solutions $\mu_1 < \mu_2 < \mu_3 < \ldots$. This implies

$$\lambda_n = \frac{\mu_n^2}{l^2}, \qquad X_n(x) = \sin\frac{\mu_n x}{l} \qquad (n = 1, 2, \ldots)$$

In this case the eigenfunctions also form a complete orthogonal system of functions.

3. Let us consider equation (2) describing the process of forced oscillations of a string under the simplest boundary conditions (16) and initial conditions (17). (By the way, the problem can be treated similarly in the case of boundary conditions of other types.) Here
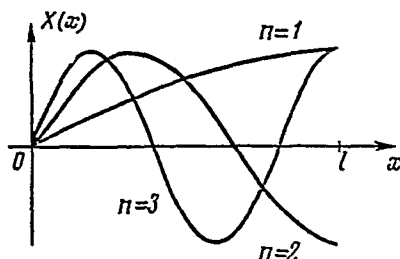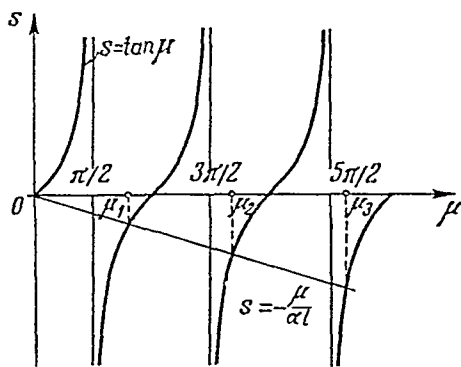
Fig. 355   Fig. 356

the first stage is to determine the set of the eigenfunctions of the corresponding homogeneous problem. This was performed in Sec. 4 and we can therefore apply the results obtained there. After that we expand the functions $u(x,)$ and $f(x,)$ into series in the eigenfunctions for any fixed value of and obtain

$$u(x, t) = \sum_{n=1}^{\infty} T_n(t) \sin\frac{n\pi x}{l} \tag{29}$$

and

$$f(x, t) = \sum_{n=1}^{\infty} H_n(t) \sin\frac{n\pi x}{l} \tag{30}$$

The coefficients $H_n(t)$ are immediately found as the Fourier coefficients of the given function $f(x, t)$:

$$H_n(t) = \frac{2}{l} \int_0^l f(x, t) \sin\frac{n\pi x}{l} \, dx$$

But the coefficients $T_n(t)$ are unknown here; they are the sought-for quantities of our problem.

Substituting (29) and (30) into equation (2) we obtain

$$\sum_{n=1}^{\infty} T_n''(t) \sin \frac{n\pi x}{l} = a^2 \sum_{n=1}^{\infty} T_n(t) \cdot \left(\frac{n\pi}{l}\right)^2 \left(-\sin \frac{n\pi x}{l}\right) +$$
$$+ \sum_{n=1}^{\infty} H_n(t) \sin \frac{n\pi x}{l}$$

Now, equating the coefficients in similar eigenfunctions we find

$$T_n'' + \left(\frac{an\pi}{l}\right)^2 T_n = H_n(t). \quad (0 \leqslant t < \infty) \tag{31}$$

Besides, if we substitute $t = 0$ into (29) then, according to initial conditions (17), its left-hand side must be equal to $\varphi(x)$. Consequently, the values $T_n(0)$ of the functions $T_n(t)$ are equal to the coefficients of the expansion of the given function $\varphi(x)$ into a series in the eigenfunctions. This enables us to easily find these values:

$$T_n(0) = \frac{2}{l} \int_0^l \varphi(x) \sin \frac{n\pi x}{l}\, dx \tag{32}$$

Similarly, differentiating equality (22) with respect to   and substituting $t = 0$ after that, we obtain

$$T_n'(0) = \frac{2}{l} \int_0^l \psi(x) \sin \frac{n\pi x}{l}\, dx \tag{33}$$

Thus, to find $T_n(t)$ we must take initial conditions (32), (33) and solve the ordinary non-homogeneous linear differential equation with constant coefficients (31). The solution of the equation can be easily found by means of the method of variation of arbitrary constants [see Sec. XV.15 and, in particular, the solution of equation (XV.87)]. Substituting the coefficients thus found into (29) we obtain the sought-for solution.

4. Let us now take a problem in which not only the equation but also the boundary conditions are non-homogeneous. For instance, let the problem be of the form

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} + f(x,\, t) \quad (0 \leqslant x \leqslant l, \quad 0 \leqslant t < \infty) \tag{34}$$

$$u|_{x=0} = \chi_1(t), \quad u|_{x=l} = \chi_2(t) \quad (0 \leqslant t < \infty) \tag{35}$$

$$u|_{t=0} = \varphi(x), \quad \frac{\partial u}{\partial t}\Big|_{t=0} = \psi(x) \quad (0 \leqslant x \leqslant l) \tag{36}$$

This problem can be reduced to a problem of the type considered above. For this purpose we take an arbitrary function $g(x,\, t)$ satisfying the boundary conditions, i.e. conditions (35). For example,

we can put $g(x, t) = \chi_1(t) + \frac{x}{l}[\chi_2(t) - \chi_1(t)]$ or choose $g(x. t)$ in any other way. Then we replace the sought-for function by means of the formula

$$u(x, t) = g(x, t) + U(x, t) \qquad (37)$$

where $U(x, t)$ is the new unknown function. To derive the differential equation and the corresponding subsidiary conditions for $U(x, t)$ we must substitute expression (37) into all the equalities (34)-(36). This results in

$$\frac{\partial^2 U}{\partial t^2} = a^2 \frac{\partial^2 U}{\partial x^2} + \left[f(x, t) + a^2 \frac{\partial^2 g}{\partial x^2} - \frac{\partial^2 g}{\partial t^2} \right] = a^2 \frac{\partial^2 U}{\partial x^2} + F(x, t)$$

$$U|_{x=0} = \chi_1(t) - g|_{x=0} = 0, \quad U|_{x=l} = \chi_2(t) - g|_{x=l} = 0 \qquad (38)$$

$$U|_{t=0} = [\varphi(x) - g|_{t=0}] = \Phi(x), \quad \frac{\partial U}{\partial t}\Big|_{t=0} = \left[\psi(x) - \frac{\partial g}{\partial t}\Big|_{t=0} \right] = \Psi(x)$$

where F, Φ and Ψ designate the expressions in the square brackets (let the reader perform all the calculations). When deducing equalities (38) we have taken into account that the function $g(x, t)$ satisfies boundary conditions (35). Thus, to find $U(x, t)$ we must now solve a problem which is completely analogous to the problem solved above, the only difference between them lying in the notation and in the particular form of the functions f, φ and ψ. After the function U has been determined we obtain the solution of the original problem by means of formula (37).

Bibliography

(STARRED ITEMS ARE IN RUSSIAN)

1. Arley, N., Buch, K. R., *Introduction to the Theory of Probability and Statistics*, New York, 1950
2*. Baranenkov, I. S., *Problems in Mathematical Analysis for Technical Colleges*, edited by B. P. Demidovich, 5th ed., Nauka, 1966
3*. Berezin, I. S., Zhidkov, N. P., *Methods of Calculations*, V. 1, Nauka, 1966, V. 2, Fizmatgiz, 1959
4*. Berman, G. N., *Collection of Problems in Mathematical Analysis*, Nauka, 1965
5*. Bermant, A. F., Aramanovich, I. G., *A Brief Course in Mathematical Analysis for Technical Colleges*, 4th ed., Nauka, 1966
6. Booth, A. D., *Numerical Methods*, London, 1955
7*. Bronstein, I. N., Semendyaev, K. A., *Reference Book in Mathematics for Engineers and Students of Technical Colleges*, 10th ed., Nauka, 1965
8*. Brudno, A. L., *Introduction to Programming*, Nauka, 1965
9. Collatz, L., *Numerische Behandlungen von Differentialgleichungen*, Berlin, 1951
10*. Demidovich, B. P., Maron, I. A., *Fundamentals of Computational Mathematics*, 3rd ed., Nauka, 1966
11*. Demidovich, B. P., Maron, I. A., Shuvalova, E. Z., *Numerical Methods of Analysis. Approximating Functions, Differential and Integral Equations*, 2nd ed., Nauka, 1963
12*. Eterman, I. I., *Analogue Computers*, Mashgiz, 1957
13*. Faddeyev, D. K., Faddeyeva, V. N., *Computational Methods of Linear Algebra*, 2nd ed., Nauka, 1963
14*. Fikhtengolts, G. M., *Course in Differential and Integral Calculus*, Vols. I–III, Nauka, 1966
15*. Fuks, B. A., Shabat, B. V., *Functions of a Complex Variable and Some of Their Applications*, 3rd ed., Nauka, 1964
16*. Gelfand, I. M., *Lectures on Linear Algebra*, 3rd ed., Nauka, 1966
17. Gnedenko, B. V., *The Theory of Probability*, Mir Publishers, 1969
18*. Gnedenko, B. V., Korolyuk, V. S., Yushchenko, E. L., *Elements of Programming*, Fizmatgiz, 1963
19*. Gradstein, I. S., Ryzhik, I. M., *Tables of Integrals, Sums, Series and Products*, 4th ed., Nauka, 1963
20*. Guter, R. S., Arlazarov, V. L., Uskov, A. V., *Reference Book in Programming*, Nauka, 1965
21*. Guter, R. S., Ovchinsky, B. V., Reznikovsky, P. T., *Programming and Computational Mathematics*, Nauka, 1965
22. Hausholder, A. S., *Principles of Numerical Analysis*, New York, 1953

23. Jahnke, E., Emde, F., *Tables of Functions with Formulas and Curves*, New York, 1945
24. Kamke, E., *Differentialgleichungen, Lösungsmethoden und Lösungen*, I., *Gewönliche Differentialgleichungen*, 6 Verbesserte Auflage, Leipzig, 1959
25*. Kitov, A. I., Krinitsky, N. A., *Electronic Digital Computers and Programming*, 2nd ed., Fizmatgiz, 1961
26. Kletenik, D., *Problems in Analytic Geometry*, Mir Publishers, 1969
27*. Krinitsky, N. A., Mironov, G. A., Frolov, G. D., *Programming*, 2nd ed., Nauka, 1966
28*. Krylov, A. N., *Lectures on Approximate Calculations*, Gostekhizdat, 1954
29*. Lebedev, N. N., *Special Functions and Their Applications*, 2nd ed., Nauka, 1963
30. Levine, L., *Methods for Solving Engineering Problems Using Analogue Computer*, New York, 1961
31*. Malkin, I. G., *Theory of Stability of Motion*, 2nd ed., Nauka, 1966
32. McCracken, D. D., *Digital Computer Programming*, New York, 1957
33. Milne, W. E., *Numerical Calculus (Approximations, Interpolation, Finite Differences, Numerical Integration and Curve Fitting)*, Princeton, 1949
34. Milne, W., Bateman, H., Bennet, A., *Numerical Integration of Differential Equations*, New York, 1956
35*. Mishina, A. P., Proskuryakov, I. V., *Higher Algebra ("Reference Books in Mathematics")*, 2nd ed., Nauka, 1965
36*. Panov, D. Yu., *Slide Rule*, 18th ed., Nauka, 1966
37. Piskunov, N., *Differential and Integral Calculus*, Mir Publishers, 1969
38*. *Practical Work in Mathematics*, edited by G. N. Polozhy, Fizmatgiz, 1960
39. *Problems in Mathematical Analysis*, edited by B. P. Demidovich, Mir Publishers, 2nd printing, 1966
40*. Romanovsky, P. I., *Fourier Series. Theory of Field. Analytical and Special Functions. Laplace Transformation*, 4th ed., Nauka, 1964
41*. Rumshisky, L. Z., *Elements of Probability Theory*, 2nd ed., Nauka, 1963
42. Salvadory, M. G., Baron, M. L., *Numerical Methods in Engineering*, New York, 1961
43*. Semendyaev, K. A., *Slide Rule. A Brief Guide Book*, 8th ed., Fizmatgiz, 1957
44*. Smirnov, V. I., *Course in Higher Mathematics*, Vols. 1, 2, Nauka, 1965, V. 3, Part 1, 7th ed., Fizmatgiz, 1956
45*. Smirnov, N V., Dunin-Barkovsky, I. V., *Course in Probability Theory and Mathematical Statistics for Technical Applications*, Fizmatgiz, 1959
46*. Smolyansky, M. L., *Tables of Indefinite Integrals*, 4th ed., Nauka, 1967
47*. Tsuberbiller, O. N., *Problems in Analytic Geometry*, 28th ed., Nauka, 1966
48*. Ventsel, E. S., Ovcharov, L. A., *Probability Theory*, Nauka, 1969
49. Yefimov, N. V., *A Brief Course in Analytic Geometry*, Mir Publishers, 1964
50*. Zagursky, V. L., *Reference Book on Numerical Methods of Solving Algebraic and Transcendental Equations*, Fizmatgiz, 1960
51*. Zeldovich, Ya. B., Myškis, A. D., *Elements of Applied Mathematics*, Nauka, 1965

Name Index

Subject Index

List of Symbols

Constants

e 68; β_n 677; B_n 678; $\begin{pmatrix} a \\ k \end{pmatrix}$ 156, 165; C 650.

Functions (general notation)

$f(x)$ 40; $y\,|_{x=a}$ 40; $f(x, y)$ 41; $z\,\big|_{\substack{x=a \\ y=b}}$ 41;

$f(\varphi(x))$ 42; $y(x)$ 42.

Functions (special notation)

$\ln x$ 69; $\exp x$ 70; $\sinh x$ 70; $\cosh x$ 70; $\tanh x$ 70; $\sinh^{-1} x$, $\cosh^{-1} x$, $\tanh^{-1} x$ 71; $\operatorname{Ln} x$ 267; $\operatorname{Erf} x$ 448; $C(x)$, $S(x)$, $Ei(x)$, $Si(x)$, $Ci(x)$ 449; $\Gamma(p)$ 469; $B(p, q)$ 471; $\delta(x)$ 489; $e(x)$ 491; $G(x, \xi)$ 492; $J_p(x)$ 568; $Y_p(x)$, $N_p(x)$ 569; $\zeta(p)$ 655; $P_n(x)$ 688, 689; $T_n(x)$ 709; $\Phi(t)$ 737, 753, 754.

Theory of Limits

$\ll$, $\gg$ 71, 121; ∞ 112; $\lim x \to a$ 113; $\underline{\lim}$, $\overline{\lim}$ 115; $\alpha \sim \beta$ 121; o, O 121, 125; min, max 131; sup, inf 170; $f(x) \sim a_0 + \dfrac{a_1}{x} + \ldots$ 686.

Finite Differences

Δ, Δx 60, 125; Δ_h 192; Δ_h^n 193, 194; $\Delta_x u$ 294.

Differential Calculus

y' 136; dy 149, 296; $\dfrac{dy}{dx}$ 150; $\dot{x}$ 152; y'', $y^{(n)}$ 155, 156; $\ddot{x}$, $d^2 y$, $d^n y$, $\dfrac{d^n y}{dx^n}$ 156, 157; u'_x 294; $\partial_x u$, $\dfrac{\partial u}{\partial x}$ 295; $\dfrac{D(f, \varphi)}{D(u, v)}$, u''_{xx}, $\partial_{xx} u$, $\dfrac{\partial^2 u}{\partial x^2}$ 303; l_x, l_y 61, 62, 138.

Vectors

$\mathbf{a}$, $\overrightarrow{AB}$, $|\mathbf{a}|$ 212; 0 215; $\text{proj}_l\,\mathbf{a}$ 219; $\mathbf{a}\cdot\mathbf{b}$, $(\mathbf{a},\,\mathbf{b})$ 220, 244; $\mathbf{r}$, $\mathbf{e}$ 221, 222; S 228; $\mathbf{a}\times\mathbf{b}$, $[\mathbf{a},\,\mathbf{b}]$ 230; E_n 239; Z_n 245.

Complex Numbers

Re, Im 259; $|z|$, $\arg z$, $\text{Arg}\,z$ 260; z^*, $\bar{z}$ 263.

Matrices

$\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ 200; $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $(a_{ij})_{mn}$, A 329; A^* 331, 333;

diag (a, b, c), I 330; $\det A$ 331; rank 337.

Field Theory

$\frac{\partial u}{\partial l}$ 365; $\text{grad}\,u$, $\text{grad}_l\,u$ 366; $\text{div}\,A$ 630; $\text{rot}\,A$, $\text{rot}_n\,A$ 636.

Integral Calculus

$\int f(x)\,dx$ 394; $\int_{\alpha}^{\beta} f(x)\,dx$ 420; $F(x)\Big|_{\alpha}^{\beta}$ 425, 426; $\bar{f}$ 434, 589;

$\oint$ 484, 629; $\int_{(\Omega)} u\,d\Omega$ 587; $\iint_{(\Omega)}$ 599.

Probability Theory

$P\{A\}$ 723; $P\{A\,|\,B\}$ 727; $\int$ 733; $\bar{\xi}$, $M\{\xi\}$,

$M\xi$ 741; $D\xi$, $D\{\xi\}$ 744; $r_{\xi,\,\eta}$ 748.

Some Other Symbols

$[M]$ 25; $\approx$ 33, 37; $\equiv$ 50; $\sum$ 117, 118; $\in$ 237;

$\bar{\in}$ 237; (R) 237; L_2 707.